SCIENTIFIC AMERICAN

FEBRUARY 1990
$2.95

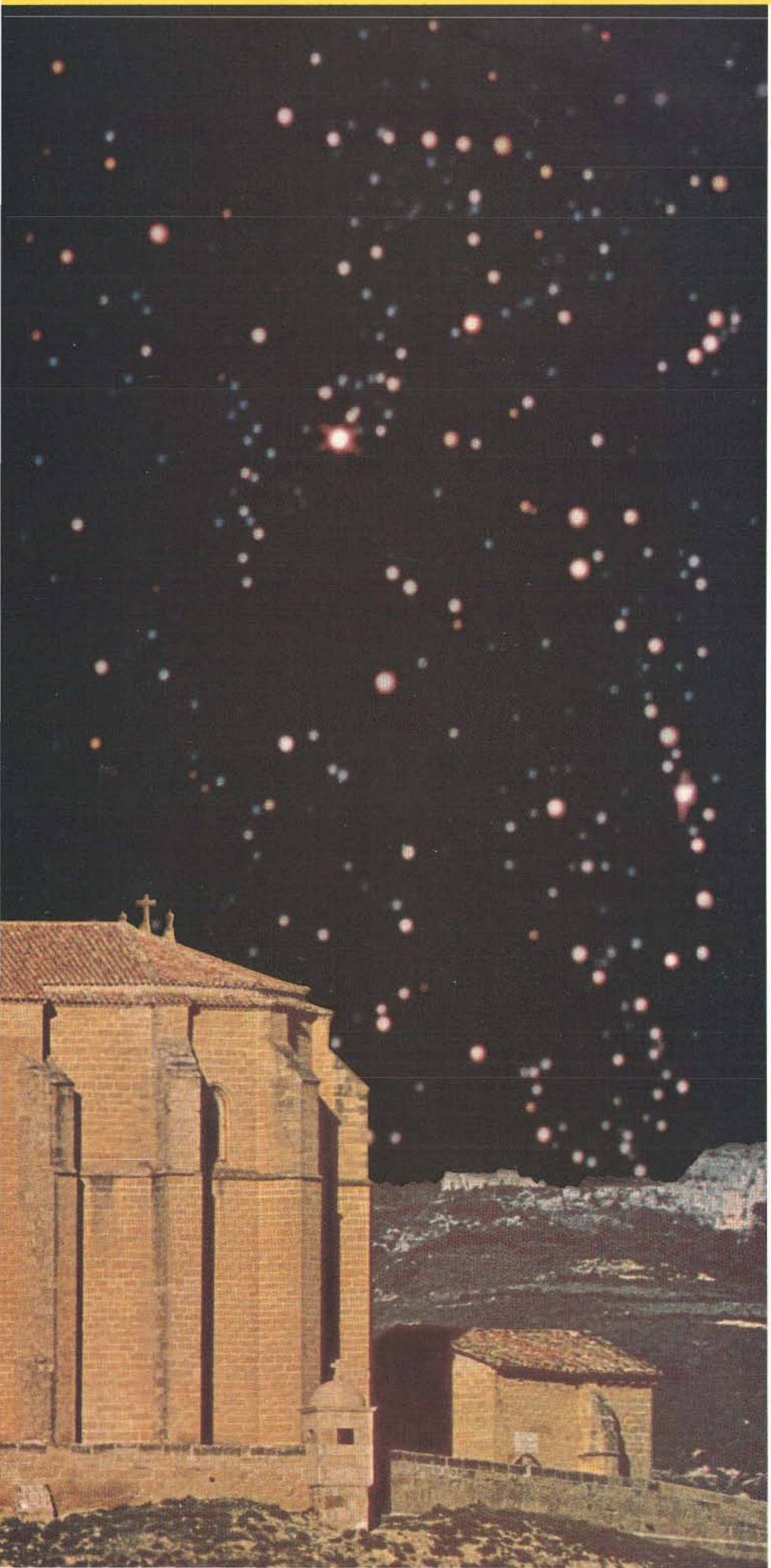
Why morphine taken for pain is not addictive.

Gallium arsenide: a dynamic technology comes of age.

Altruism in—would you believe it?—vampire bats.

Variable inferno: *the churning sun grows brighter and more turbulent as it approaches the peak of its 11-year activity cycle.*

To follow the stars, as the Pilgrims did, is to feel the true spirituality of Spain.



Heavenly Spain.

This church is on the route to Compostela which, literally translated, means "The Field of the Stars".

Guided by those stars, the medieval Pilgrims used to worship at churches such as this one, when on their way to visit the tomb of St James at Santiago de Compostela.

The area is typified by such important places of historical interest as Pamplona, Logroño, Burgos, Leon and Lugo.

They are rich in cathedrals, churches, monasteries and city walls, built thousands of years ago.

Enjoy a memorable meal in one of the typical Gallician restaurants –the seafood and fish are out of this world.

It's all part of the heavenly experience.

All this and sunshine too.

Spain. Everything under the sun.

ESPAÑA

SCIENTIFIC AMERICAN

February 1990 Volume 262 Number 2

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Copyright © 1990 by Scientific American, Inc. All rights reserved. Printed in the U.S.A. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Subscription rates: one year $27, two years $48, three years $66 (outside U.S. and possessions add $11 per year for postage). Subscription inquiries: U.S. only 800-333-1199; other 515-247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51593.

DEPARTMENTS

THE COVER painting, based on a photograph made (in the light emitted by hydrogen) at the National Solar Observatory, shows the fierce turmoil typical of the sun's lower atmosphere during times of high solar activity. The number of bright flares and dark filaments rises and falls in an 11-year cycle linked to changes in the sun's magnetic field (see "The Variable Sun," by Peter V. Foukal, page 26). The cycle also alters the sun's output of radiation and charged particles, which can affect conditions on the earth.

THE ILLUSTRATIONS

Cover painting by Ian Worpole

Page	Source
20-21	Patricia J. Wynne
23	Carol Donner
24	Linda Bartlett
27	National Solar Observatory, Sacramento Peak, New Mexico
28	Ian Worpole
29	Naval Research Laboratory, Washington, D.C. (*top*), National Solar Observatory, Kitt Peak, Arizona (*bottom*)
30-32	Ian Worpole
33	Peter V. Foukal (*left*), National Aeronautics and Space Administration (*right*)
34	Quesada/Burke, cast courtesy of John Lehr
35	Yoichiro Kawaguchi
36-37	Carol Donner, redrawn from *A Textbook of Histology*, by Don W. Fawcett, W. B. Saunders Co., © 1986 (*top*), Joseph Mietus, Beth Israel Hospital, Boston, and Edward Bell (*bottom right*)
38	Burton Sobel, Washington University School of Medicine
39	Carol Donner
40	Edward Bell and Joseph Mietus
41	Carla J. Shatz, Stanford University School of Medicine
43	Quesada/Burke
44-46	George Retseck
47	Johnny Johnson
50	George Retseck
57	Randall M. Feenstra and Joseph A. Stroscio, IBM Thomas J. Watson Research Center
58-59	George V. Kelvin
60	International Business Machines
62	George V. Kelvin
63	Edward Bell
65	Gunter Ziesler, Peter Arnold, Inc.
66	M. W. Larson, Bruce Coleman, Inc. (*top*), Patricia J. Wynne (*bottom*)
67	Gerald S. Wilkinson
68-69	Patricia J. Wynne
70	Merlin D. Tuttle
73	*Novgorod, Monuments of Art and Architecture*, Aurora Publishers, ©1984
74	Tomo Narashima
75-78	Valentin L. Yanin
79	Tomo Narashima
81	Andrew Christie
82	Casa Editrice Giusti di Becocci, Firenze
83-84	Andrew Christie
85	Paul Wallich
86	Jearl Walker
87-89	Michael Goodman

SCIENTIFIC AMERICAN®

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Armand Schwab, Jr., *Managing Editor;* Timothy Appenzeller, Laurie Burnham, *Associate Editors;* Timothy M. Beardsley; Elizabeth Corcoran; John Horgan; June Kinoshita; Philip Morrison, *Book Editor;* Corey S. Powell; John Rennie; Philip E. Ross; Ricki L. Rusting; Russell Ruthen; Paul Wallich; Karen Wright

ART: Samuel L. Howard, *Art Director;* Edward Bell, Joan Starwood, *Associate Art Directors;* Johnny Johnson

COPY: Maria-Christina Keller, *Copy Chief;* Nancy L. Freireich; Michele S. Moise; Philip M. Yam

PRODUCTION: Richard Sasso, *Vice President Production and Distribution; Managers:* Carol Eisler, *Manufacturing and Distribution;* Carol Hansen, *Electronic Composition;* Leo J. Petruzzi, *Manufacturing and Makeup;* Carol Albert; Madelyn Keyes; William Sherman

CIRCULATION: Bob Bruno, *Circulation Director;* Lorraine Terlecki, *Business Manager;* Cary Zel, *Promotion Manager*

ADVERTISING: Robert F. Gregory, *Advertising Director.* OFFICES: NEW YORK: Peter Fisch; John Grant; Meryle Lowenthal; William Lieberman, Inc. CHICAGO: 333 N. Michigan Avenue, Chicago, IL 60601; Patrick Bachler, *Advertising Manager;* Litt Clark, *Midwest Manager.* DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; William F. Moore, *Advertising Manager;* Edward A. Bartley, *Detroit Manager.* WEST COAST: 1650 Veteran Avenue, Suite 101, Los Angeles, CA 90024; Kate Dobson, *Advertising Manager;* Joan Berend, San Francisco. ATLANTA, BOCA RATON: Quenzer/Stites. CANADA: Fenn Company, Inc. DALLAS: Griffith Group.

ADVERTISING SERVICES: Laura Salant, *Sales Services Director;* Diane Greenberg, *Promotion Manager;* Ethel D. Little, *Advertising Coordinator*

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager,* London; GWP, Düsseldorf. HONG KONG/SOUTHEAST ASIA: C. Cheney & Associates. SEOUL: Biscom, Inc. SINGAPORE: Cheney Tan Associates. TOKYO: Nikkei International Ltd.

BUSINESS MANAGER: Marie D'Alessandro

PUBLISHER: John J. Moeling, Jr.

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue
New York, NY 10017
(212) 754-0550

PRESIDENT AND CHIEF EXECUTIVE OFFICER: Claus-Gerhard Firchow

EXECUTIVE COMMITTEE: Claus-G. Firchow; *Executive Vice President and Chief Financial Officer,* R. Vincent Barger; *Vice Presidents:* Linda Chaput, Jonathan Piel, Carol Snow

CHAIRMAN OF THE BOARD: Georg-Dieter von Holtzbrinck

CHAIRMAN EMERITUS: Gerard Piel

LETTERS

To the Editors:

Your "Science and the Citizen" item "Land-Locked" [SCIENTIFIC AMERICAN, October, 1989] echoes Fred C. Iklé's question "Why on earth, heaven and hell do we still want the land-based ICBM?" and then turns to Iklé, Barry Blechman, Marshall Brement and Robert S. McNamara for answers. You would do better to consider a few salient facts than to consult ideologues.

There are six simple reasons why we still want the land-based ICBM. First, ICBM's are far less expensive than bomber aircraft and submarine-based missiles. Whatever one's measure—life-cycle cost, total force cost, cost per delivered yield, cost per alert megaton—the most cost-effective weapon in our strategic arsenal is the ICBM.

Second, ICBM's perform the fundamental strategic-attack mission much better than aircraft or submarines. The ability of bombers to reach defended targets is relatively limited; the limited yield and accuracy of sea-based missiles precludes their use against time-urgent hard targets. ICBM's alone can promptly deliver large yields with pinpoint accuracy; other strategic systems merely supplement this fundamental mission.

Third, ICBM's are much more reliable—and consequently available—than bombers and submarines. At any given moment, virtually the entire ICBM force is on alert, ready for immediate launch. At that same moment, an expensive fraction of our bombers and submarines is sitting in hangars and shipyards or is otherwise out of commission.

Fourth, ICBM's do not depend on fragile communications links to perform their mission, unlike bombers and submarines. One consequence is that ICBM's cannot be defeated or misled by simple electronic countermeasures.

Fifth, land-based ICBM's present an extremely high deterrent to a would-be attacker. An enemy can destroy our bombers and submarines in the air and oceans, even without employing nuclear weapons, and expect to suffer nothing worse in return. To contemplate an attack on our ICBM's, however, an enemy must be prepared to launch an extensive strategic nuclear attack on the American heartland and be willing to suffer massive retaliatory destruction.

Sixth, a strategic triad that includes ICBM's—especially mobile ICBM's—is inherently a better deterrent, now and in the long term, than a dyad of bombers and submarines. The triad's deterrence derives, first, from the operational impossibility of mounting concurrent surprise attacks on bombers, submarines and two kinds of ICBM's (silo-based and mobile). Equally important, the most convincing threat of retaliation resides in the triad's ICBM's, the weapons that are the least susceptible to neutralization by new defensive technologies. Although air defense against bombers is expensive, the technology is by no means exotic. Similarly, the primary obstacles to effective antisubmarine warfare—at least in the U.S.—are budgetary and bureaucratic. However, as the Strategic Defense Initiative demonstrates, effective defense against ICBM's is an enormous challenge, technical as well as economic—one that is not likely to be met before manned bombers and submarines are rendered obsolete.

JACK H. HARRIS
President,
Center for National
Program Evaluation
Reston, Va.

To the Editors:

I trust you can understand my surprise, as one of nearly three million inhabitants of a metropolitan area on the north shore of Lake Ontario, to discover from the cover of your September, 1989, single-topic issue that I live in the dark. How does this square with the fact that, per capita, Canadians are the most profligate consumers of energy in the world? What's more, it appears that all eight to nine million fellow Ontarians who live in the area from Windsor east to the Quebec border and north to Sudbury also live in total darkness. Sudbury, a city of approximately 90,000, is the apex of this triangle and seems to be the only urban area with lights, in spite of the fact that this triangle is the most heavily industrialized part of Canada!

There is a Canadian game, which contributes greatly to national unity, called "Let's All Hate Toronto and Forget It Ever Existed," but Americans are not qualified to play.

C. C. BARNES
Toronto, Canada

To the Editors:

I applaud the in-depth coverage in your September, 1989, issue of the multifaceted problems facing the earth's inhabitants today. I appreciate your authors' willingness to spell out some of the changes in direction that humans must make to achieve sustainable development. Yet I am frustrated by a serious factual error in one of the otherwise excellent articles: "Strategies for Energy Use," by John H. Gibbons, Peter D. Blair and Holly L. Gwin.

At the top of page 88 you present a graph illustrating world energy consumption. The graph erroneously shows a decline in total world consumption of roughly 5 percent between 1975 and 1985. In contrast, figures from the U.S. Energy Information Administration indicate that world energy consumption grew 24 percent in that decade. Industry sources such as the Chevron Corporation and the British Petroleum Company confirm the EIA figure.

World energy consumption doubled between 1965 and 1988 and is still on the rise. If the annual growth experienced during the past five years is allowed to continue, world energy consumption will double again in 20 years. As your authors explain, this could have disastrous consequences for the environment in which our children will be living.

STEVE ANDREWS
Denver, Colo.

To the Editors:

While reading "Authenticating Ancient Marble Sculpture," by Stanley V. Margolis [SCIENTIFIC AMERICAN, June, 1989], I began to assume that the Greek kouros on page 80 was fraudulent, because the right testicle is shown as hanging lower than the left. People familiar with the human anatomy know that the left testicle almost always hangs lower than the right; the sculptures of David by both Donatello and Michelangelo are illustrative. Is this anatomical fact a comparatively recent discovery, or was the picture in the article printed in reverse?

L. M. KLEVAY
Grand Forks, N.D.

The photograph was indeed reversed.
—THE EDITORS

Every month we receive hundreds of letters from our readers. We and our authors thank you for sharing your thoughts with us—and ask for your forbearance. The sheer number of letters makes it impossible for us to answer more than a fraction of them.

50 AND 100 YEARS AGO

FEBRUARY, 1940: "AIR, WATER, COAL = HOSIERY. A plant for the commercial manufacture of nylon yarn, erected by the du Pont Company at Seaford, Delaware, went into production on December 15 last. Nylon hosiery will be made by a number of nationally known hosiery manufacturers, and it is anticipated that nylon hosiery for both men and women will be put on the general market by late spring or early summer of 1940. The manufacturers say that the word 'nylon' has no significant derivation; that it was selected as a generic name because it is non-technical and is easy to pronounce; its individual letters do not stand for anything."

"In the laboratory of Professor Harold E. Edgerton, at the Massachusetts Institute of Technology, one can see time not only arrested but practically handcuffed by the modern stroboscope. A bright light from a black box is flashed onto a stream of water which in ordinary light appears to be continuous. In a moment the stream becomes a series of drops that stand still—queerly misshapen little jewels poised in midair. A turn of a knob, and they seem to climb back up into the faucet, defying the law of gravity. The stroboscope is not merely a toy for the creation of uncanny hallucinations. In technology and industry the stroboscope is rapidly coming into its own. Its commonest uses are the observation of rapidly revolving machinery and the measurement of the rate of revolution."

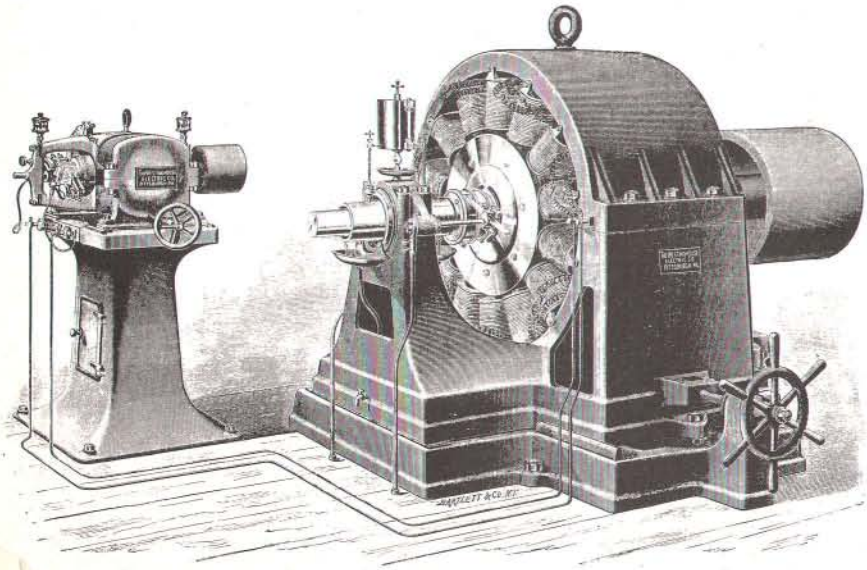
"A Douglas DC-3 transport airplane, operated by United Airlines, equipped with supercharged Pratt & Whitney engines, recently and unwittingly broke the world's altitude record for the type of aircraft by flying to a height of 28,900 feet. It carried two scientists of the University of Chicago who were engaged in photographing mesotrons, the heavy radioactive components of cosmic rays. The flight was part of the cosmic-ray investigations being carried on by Dr. Arthur H. Compton, the Nobel prize winner, who maintained contact with the airplane by radio."

"What is said to be the first basic improvement in combs in 4,000 years, a new one, called Komatic, has been placed on the market. In this comb the backbone has been pushed completely to one side so that the teeth slots go all the way to the back edge. Its self-cleaning characteristic is said to be due to the fact that this side backbone gives a sharp edge on one side and there are no pockets in which hairs or lint may collect."

FEBRUARY, 1890: "When it was stated some weeks since in the newspapers that the building of a milk pipe line from a point in New York State to New York City was projected there was a rather general smile, and the matter was treated as a joke. The projectors were, however, it seems, in sober earnest. A company with a capital of $500,000 has been formed for the purpose of constructing such a line. The proposed method of forwarding the milk is in cylindrical tin cans surrounded and propelled by water, and the promoters of the scheme assert that the time of transportation for a distance of 100 miles will not exceed an hour."

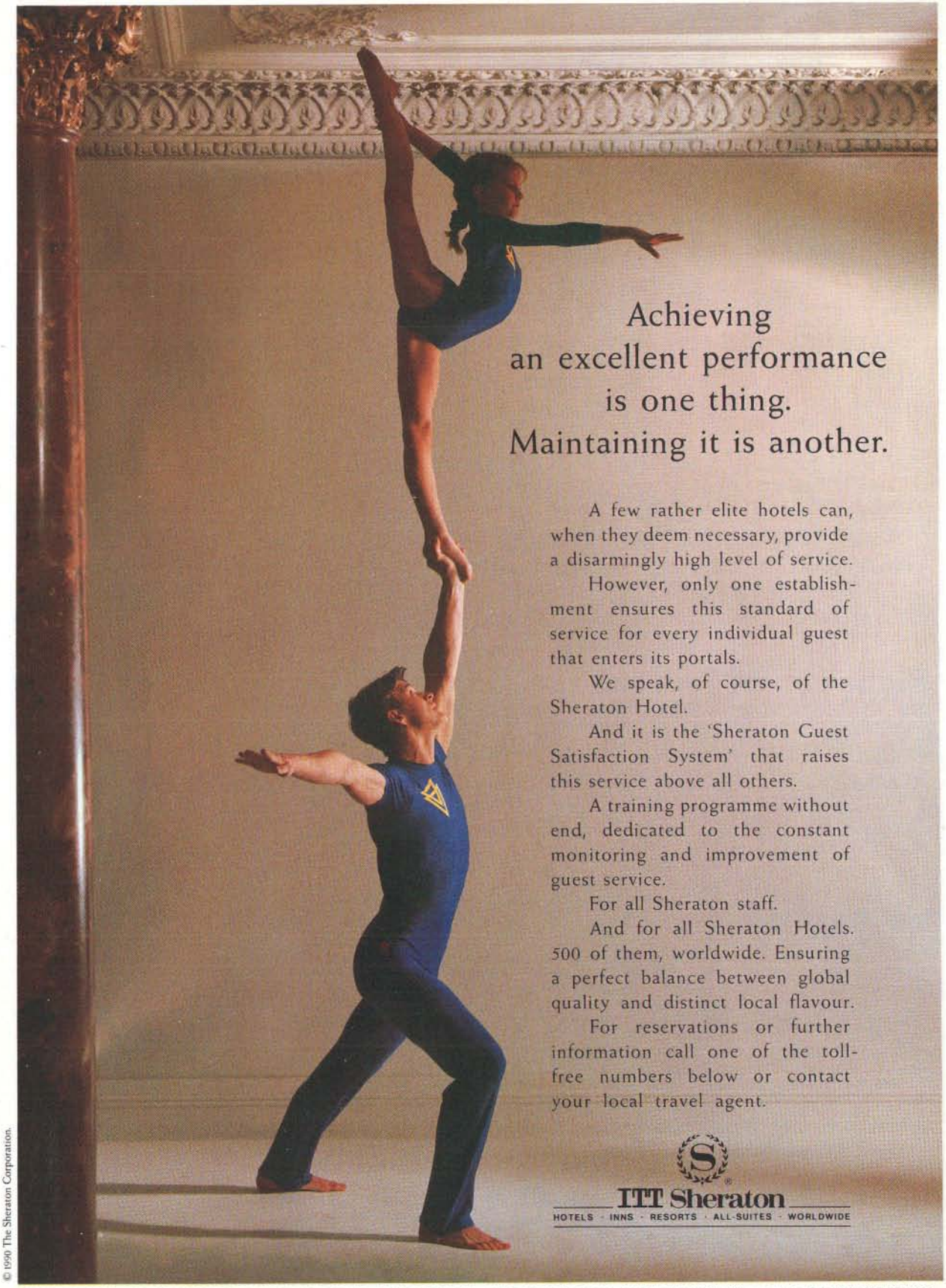
"Few Londoners are aware that there are now under the streets of the metropolis forty miles of pipes charged with water at a pressure of 750 pounds per square inch. These are the mains of the London Hydraulic Power Company. Compressed air has been largely used for transmitting power in England, notably in Birmingham, on the Continent, and in the United States; and electricians are working hard with a view to the introduction of electricity as the agent. But in London the system of hydraulic power is virtually having its own way. This power is supplied direct to lifts, presses, and other purposes of a similar character without the use of any engine or power-producing machinery. It is available day and night and on Sundays all the year round. There it is, to be had by the turning of a tap."

"In the successful development of central station electrical distribution, electrical engineers have been confronted by a dilemma. Considerations of safety to life require the use of a low tension circuit when there is any possibility of human beings touching the wires. Considerations of economy in the use of copper require the use of a high tension circuit to enable small wire to be used. The dilemma has been solved by the alternating current converter system. The Westinghouse system, an exponent of the alternating current installation, includes a complete series of appliances, from central station dynamo to the customer's individual converter and meter. The dynamo, illustrated here, includes a field excited by a small direct current generator, shown in the background upon a pedestal. The field contains an even number of cores, placed radially as shown, so wound that the cores terminate alternately in north and in south poles."



Alternating current: the Westinghouse dynamo and exciter

· AT SHERATON LITTLE THINGS MEAN A LOT · AT SHERATON LITTLE THINGS MEAN A LOT · AT SHERATON LITTLE THINGS MEAN A LOT ·



Achieving
an excellent performance
is one thing.
Maintaining it is another.

A few rather elite hotels can, when they deem necessary, provide a disarmingly high level of service.

However, only one establishment ensures this standard of service for every individual guest that enters its portals.

We speak, of course, of the Sheraton Hotel.

And it is the 'Sheraton Guest Satisfaction System' that raises this service above all others.

A training programme without end, dedicated to the constant monitoring and improvement of guest service.

For all Sheraton staff.

And for all Sheraton Hotels. 500 of them, worldwide. Ensuring a perfect balance between global quality and distinct local flavour.

For reservations or further information call one of the toll-free numbers below or contact your local travel agent.

ITT Sheraton
HOTELS · INNS · RESORTS · ALL-SUITES · WORLDWIDE

© 1990 The Sheraton Corporation.

BELGIUM 11 35 35 · FRANCE 19 05 90 76 35 · HOLLAND 06 03 35 · ITALY 16 78 35 035 · SWEDEN 020 79 58 35 · UNITED KINGDOM 0800 35 35 35. GERMANY 0130 35 35.
BRUSSELS · CATANIA · COPENHAGEN · EDINBURGH · ESSEN · FRANKFURT · GOTHENBURG · ISTANBUL · JERUSALEM · LIMASSOL · LISBON · LONDON, (BELGRAVIA, HEATHROW, PARK TOWER, SKYLINE) · LUXEMBOURG · MADEIRA · MALMO · MUNICH · PADUA · PALMA DE MAJORCA · PORTO · ROME · SALZBURG
SOFIA · STOCKHOLM · TEL AVIV · ZURICH · BISSAU · COTONOU · DJIBOUTI · HARARE · KAMPALA · LAGOS · LIBREVILLE · LUBUMBASHI · SEYCHELLES

AT SHERATON LITTLE THINGS MEAN A LOT · AT SHERATON LITTLE THINGS MEAN A LOT · AT SHERATON LITTLE THINGS MEAN A LOT · AT SHERATON LITTLE THINGS MEAN A LOT · AT SHERATON LITTLE THINGS MEAN A LOT · AT SHERATON LITTLE THINGS MEAN A LOT ·

· AT SHERATON LITTLE THINGS MEAN A LOT · AT SHERATON LITTLE THINGS MEAN A LOT · AT SHERATON LITTLE THINGS MEAN A LOT ·

SCIENCE AND THE CITIZEN

R.I.P. Blackbird

A legendary spy plane is brought down to earth

The passing of the Cold War offers little to lament. Who cherishes the Berlin Wall, or cruise missiles tipped with nuclear warheads? But one need not be a hawk to regret the demise of one symbol of superpower hostility: the SR-71 Blackbird, which the U.S. Air Force intends to retire this year. Created more than 25 years ago, the spy plane is still the fastest, highest-flying aircraft in existence. It makes even professional critics of military technology sound like boys in the model-airplane phase.

"Definitely one of the most beautiful airplanes ever built," asserts John E. Pike of the Federation of American Scientists. Pike recalls examining a grounded Blackbird at Wright Patterson Air Force Base in Ohio and finding that its titanium skin was "thin as a Coke can's." "I assumed something that flew that fast would have some kind of heavy plating," Pike says. "It was just so amazing."

Blackbird was conceived in the late 1950's, as U.S. officials began worrying that the U-2, a high-flying but slow jet glider, could be shot down on spy missions over the U.S.S.R. That fear was borne out in 1960 when Soviet missiles felled a U-2 flown by Francis Gary Powers. A few years later Lockheed's secret "Skunk Works" in Burbank, Calif., the same outfit that produced the U-2, hatched a plane that could fly higher and much, much faster.

Blackbird is the original "stealth" aircraft: its sleek curves offer radar few handholds, and special epoxy coatings reduce its reflectance still further. It is the first and will probably be the last plane made almost entirely of titanium; the toughness that enables the metal to withstand the great heat and pressure of supersonic flight also makes it hard to work with. Two Pratt & Whitney turboramjet engines some 40 feet long generate more than 32,000 pounds of thrust each.

The SR-71's true talents are classified: the Air Force has revealed only that it *cruises* at Mach 3—literally faster than a speeding bullet—at altitudes of about 85,000 feet. In his book *Deep Black*, William E. Burrows of New York University estimates the plane's top speed at 2,600 miles per hour and its ceiling at 105,000 feet.

Some 30 Blackbirds have been built, according to Jeffrey T. Richelson of the National Security Archives. They have spied on virtually all the postwar antagonists of the U.S., including the Soviet Union, North Korea, North Vietnam, Iran, Libya and Nicaragua. Although SR-71's have reportedly been fired on more than 1,000 times, they have never been shot down. Accidents and retirement, however, have reduced the original flock of Blackbirds to about 20, Richelson says; at any given moment perhaps half of these are ready to fly out of bases in California, England and Okinawa.

For several years the Air Force has tried to retire the remaining Blackbirds, arguing that satellites and other planes—including new versions of the old U-2—can do their job more cheaply. Indeed, everything needed to keep the SR-71 flying—spare parts, fuel and lubricants, pilot training—is exotic and extraordinarily expensive. Nevertheless, Congress has kept the SR-71 program alive until this year, when the Air Force finally got its way.

"There's nothing else that can replace it," mourns James Currie, an aide to the Senate Select Committee on Intelligence. Unlike a satellite, Currie notes, an SR-71 can be sent "where you need it, when you need it," and unlike U-2's or any other spy plane, the Blackbird "can penetrate hostile air space and get out alive." These talents would be particularly valuable in the volatile post-Cold War era, Currie says, in which localized conflicts may replace the Soviet bloc as a primary source of concern to the U.S.

Pike agrees. He points out, moreover, that the expense of keeping the SR-71's flying—which he estimates at $300 million a year—is still a fraction of the cost of a single reconnaissance satellite. The Pentagon has at least four such satellites in orbit now—three optical and one radar-based—

Expensive, leaky and "not real maneuverable," the old SR-71's may still be sorely missed

SR-71 BLACKBIRD measures 107 feet from tip to stern and 56 feet across the wings.

No other system of keeping up can compare with ours.

YOUR SYSTEM: a time-consuming, futile struggle to keep up with the information explosion.

The classic texts are convenient references—but the information they contain is obsolete before publication.

Like many physicians, you probably rely on the texts you first used in medical school. But even using the most recent editions, you find material that no longer reflects current clinical thinking—and they lack the latest information on such topics as herpes, oncogenes, AIDS, and photon imaging.

Reading stacks of journals alerts you to recent developments—but can't give you quick answers on patient management.

Struggling through the hundreds of journal pages published each month—even on only the really significant advances in the field—is arduous and memory-taxing. And it's a task that costs physicians valuable time—their most precious resource.

Review courses cover clinical advances—but, months later, do you recall the details of a new procedure or unfamiliar drug?

Seminars can also be costly and make you lose valuable time away from your practice—expenses that may amount to several thousand dollars. And, the speaker's skill often determines how much you learn.

OUR SYSTEM: a rewarding, efficient way to keep yourself up-to-date—and save hundreds of hours of your time for patient care.

A comprehensive, 2,300-page text in two loose-leaf volumes, incorporating the latest advances in medical practice as of the month you subscribe.

This superbly designed, heavily illustrated resource, called "the best written of all [the internal medicine] books" by *JAMA* (251:807, 1984), provides a practical, comprehensive description of patient care in 15 subspecialties. And, because the text is updated each month, the clinical recommendations reflect all the current findings. A practice-oriented index and bibliography of recent articles further enhance the efficiency of the text.

Each month, six to nine replacement chapters to update your text *plus* new references, an eight-page news bulletin, and a completely new index.

You'd have to read hundreds of journal pages each month—and memorize the contents—to get the same information SCIENTIFIC AMERICAN *Medicine* contains. With our updated text, you read only the information you really need. Our authors, largely from Harvard and Stanford, sort through the literature and monitor developments, incorporating the significant advances into our chapters.

At no additional cost, a 32-credit CME program, to save you valuable patient-care time and the expense of attending review courses.

Earn 32 Category 1 or prescribed credits per year with our convenient self-study patient management problems; each simulates a real-life clinical situation. Choose either the complimentary printed version or, at a modest extra charge, the disk version for your IBM® PC or PS/2, Macintosh™, or Apple® (or compatible).

☐ Yes, I'd like to try the SCIENTIFIC AMERICAN *Medicine* system. Please enter my subscription at a first-year price of US$265* plus $7 shipping for a total of US$272.

☐ Also enroll me in the CME program. I'd prefer:
- ☐ the printed version at no additional charge.
- ☐ the disk version at US$97* additional. Computer type or compatible:
  - ☐ IBM® PC (256K) or compatible (5¼") ☐ IBM® PS/2 (3½") ☐ Macintosh™ (512K)
  - ☐ Apple® IIc/IIe/IIGS ☐ Apple® II + with 80-col. card by: ☐ Apple® ☐ Videx®

☐ Enroll me in the disk CME only at US$152.* (Note model above.)

☐ Check enclosed* ☐ Bill me

☐ VISA ☐ MasterCard Exp. Date _____ Account No. _____

Name _____ Specialty _____

Address _____

City _____ State _____ Zip Code _____

9110

*Add sales tax if for CA, DC, IL, MI or NY. Allow 8 weeks for delivery. Add US$10 for shipping to Canada. IBM is a registered trademark of International Business Machines Corporation. Apple is a registered trademark of Apple Computer, Inc. Videx is a registered trademark of Videx, Inc. Macintosh™ is a trademark of McIntosh Laboratory, Inc. and is used by Apple Computer, Inc. with its express permission.

Or, call toll-free: 1-800-345-8112

Your subscription includes:

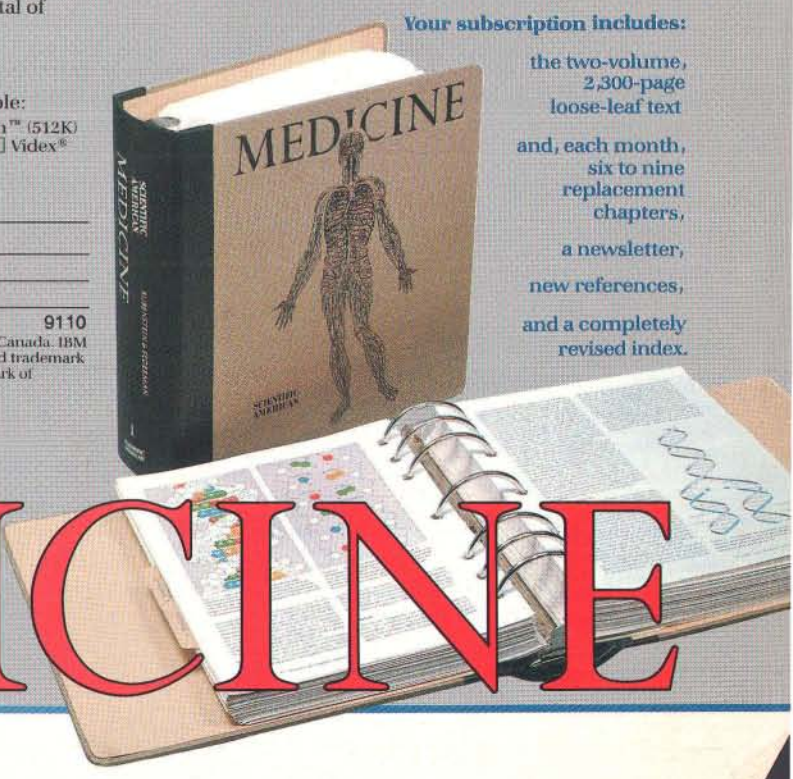
the two-volume, 2,300-page loose-leaf text

and, each month, six to nine replacement chapters,

a newsletter,

new references,

and a completely revised index.



SCIENTIFIC AMERICAN

MEDICINE

and plans to have as many as 12 by the mid-1990's, according to Pike. Because these satellites could be shot down by antisatellite weapons, he says, Blackbirds provide at the very least an invaluable backup capability.

Major Gary I. Luloff of the Strategic Air Command, who once piloted SR-71's and is now a manager of the reconnaissance program, admits that the plane has its flaws: it is "not real maneuverable"; its titanium fuselage, before heat seals the joints by making the metal expand, leaks fuel like a sieve; it gets terrible gas mileage and so must be frequently refueled by airborne tankers during missions; its pilots must wear heavy, hot astronaut suits. But when asked about the official Air Force position that other systems can do what Blackbird has done, Luloff's voice takes on an edge, as if someone had slighted an old friend: "I think it's fair to say the demise of the program was more of a budgetary one than a capability one."

Any chance that the Air Force brass will reconsider? "No," replies Major Richard M. Cole, an Air Force spokesman. "The program is terminated."

Bye-bye, Blackbird. —*John Horgan*

Aborted Research

Ideology seems to have put some medical advances on hold

In late 1987 Edward H. Oldfield and Robert J. Plunkett of the National Institute of Neurological and Communicative Disorders and Stroke saw administrative fiat terminate a promising area of their research.

They had offered experimental surgery to a patient suffering from Parkinson's disease. They would graft healthy cells from the brain of an aborted fetus into the patient's brain, in the hope that the fetal cells would compensate for cells the disease had killed. Experiments with animals had been encouraging. According to the research protocol, which had been approved by the relevant ethics committees and by the director of clinical research for the National Institutes of Health, a woman who had already decided to have an abortion, in a different hospital, would have been offered the opportunity to donate fetal tissue for the procedure. The Parkinson's patient was admitted. At the last moment, however, the director of the NIH ordered that the surgery be delayed until the Department of Health and Human Services gave its approval. The approval never came.

Shortly afterward the assistant secretary for health ordered a moratorium on federal funding for all experiments involving fetal tissue from elective abortions. The effect was to stymie fetal-tissue transplantation not only for Parkinson's disease but also for diabetes and Alzheimer's disease.

The NIH convened a panel to examine the ethical questions. The members concluded that transplantation of fetal tissue was permissible public policy, provided that safeguards were enacted to ensure that the prospect of helping a sick person would not persuade a woman to have an abortion. Notwithstanding the panel's conclusion, last November Secretary of Health and Human Services Louis W. Sullivan announced a permanent ban on federal support for transplantation of tissue from elective abortions, over the opposition of several disease foundations and the Association of American Medical Colleges.

Fetal-tissue transplantation has become the latest casualty of a war in which foes of abortion have "stifled in the country, and repressed at the federal level, research into human reproductive biology," John C. Fletcher of the University of Virginia at Charlottesville says. The repression is apparently being abetted by an administration mindful of its political debts. According to NIH researchers, senior officials there have even prevented the publication of a review of research on the abortifacient drug RU-486. And it is a commonplace among NIH workers that the political climate is discouraging workers from studying human fertility. As a result, Fletcher believes, the level of U.S. fertility research has been "less than adequate."

Officials have also applied Kafkaesque bureaucratic rules to frustrate research related to infertility, a problem that affects more than 8 percent of women of childbearing age. In 1987 Oliver H. Lowry of the Washington University School of Medicine sought an NIH grant to improve a culture medium required for in vitro fertilization. An NIH review panel gave the proposed research, which would have used excess eggs destined for disposal by IVF clinics, a near-perfect assessment. But the NIH also told Lowry that before the work could continue he would need to procure the approval of the DHHS's Ethics Advisory Board. Therein lay the catch-22: the board had ceased to exist in 1980, and the assistant secretary for health, James O. Mason, has declined to resuscitate it. Lowry still does not have his grant.

Fletcher suggests that such ideologically motivated screening of research recalls the influence of Trofim Lysenko, the geneticist whose doctrinaire views smothered biological research in the Soviet Union for many years. Certainly the effects of such control have come to extend beyond infertility and degenerative diseases. For example, Fletcher points out, research on how soon genes are expressed in the early embryo cannot be federally funded. Yet if proteins signifying genetic disease could be detected in eggs fertilized in vitro, embryos lacking such proteins could be selected for implantation. The NIH estimates that more than 100 grant applications for early-embryo research would be filed if the ethics board were in place.

The political standoff is also hampering research on fetal abnormalities and on such diagnostic procedures as chorionic villus sampling. Federal support for any such work that involves more than a strictly defined minimal risk needs the approval of a congressional ethics board; political stalemates have paralyzed the board and its advisory committee since 1985.

The government's reluctance to countenance research involving embryos and fetuses has had a chilling effect. Private foundations and even large pharmaceutical companies are becoming skittish about sponsoring such work, according to D. Eugene Redmond, Jr., of the Yale University School of Medicine. In the meantime other countries, Sweden and Canada in particular, are making headway.

The battle extends beyond the laboratory and the operating room: two reports released in December by Representative Ted Weiss's subcommittee on human resources excoriate successive administrations for attempts to censor epidemiological studies of abortion as well as for failing to support research on infertility. Weiss plans to hold hearings on the subject this year. Most workers, nevertheless, see little hope of a reprieve for either infertility studies or research on transplantation and fetal diseases in the near future. —*Tim Beardsley*

Chub Hubbub

Environmental concerns clash in California's Mohave desert

An environmental saga with a cast of thousands and an epic time scale is unfolding in California's Mohave desert. Would some ecology-conscious producer—Robert Redford, perhaps—like to scan the script?

Fade in to the close of the last ice age, as glaciers retreating from the Mohave region leave behind pockets of water inhabited by various hardy species of fish. One of these survivors is the Mohave tui chub, an olive-drab, pug-nosed minnow from four to eight inches long.

Cut to the late 1960's. The number of Mohave tui chub is dwindling in the fish's native habitats—partly as a result of interbreeding with other species of chub—and state wildlife officials seek new homes where its genealogical purity might be preserved. The officials identify several promising sites, including a marsh-ringed pond, called Lark Seep, in a corner of the China Lake Naval Weapons Center. China Lake is actually a vast, parched lake bed where the Navy tests instruments of war. Would the Navy mind sharing its base with a few harmless fish? The Navy can think of no objections, and 400 chub move into Lark Seep in 1971. A few years later, significantly, the U.S. Fish and Wildlife Service declares the chub to be an endangered species.

Everything is fine until the early 1980's, when a strange problem begins to afflict the China Lake base: too much water. It seeps into the basements of buildings and rots their foundations; it undermines roads; it creates marshes that attract snow geese, which can do a lot of damage when sucked into a jet engine. The water is highly alkaline, and some authorities worry that it might also contaminate the underlying aquifer. Apparently the water is traveling underground from some unknown source.

Officials of California's Lahontan Regional Water Quality Control Board investigate the mystery. They suspect that the renegade water is leaking from a sewage-treatment plant serving both the 6,000 residents of China Lake and the 30,000 residents of Ridgecrest, a nearby desert community. Although the Navy built the sewage plant, which consists of a filtration system and a number of evaporation ponds, Ridgecrest has operated it since the late 1970's.

In 1987 the water-quality board orders the city to plug the leaks in the sewage plant or face stiff fines. The city is happy to comply; with the help of consultants from Bechtel, Inc., the city even devises a plan that would allow it to profit from the problem, by selling treated sewage water to a nearby golf course or chemical company. The Navy, anticipating an end to its watery woes, is elated.

But there is a small catch. It seems that water from the sewage plant is the lifeblood of Lark Seep, and Lark Seep has become the world's largest repository of Mohave tui chub. The seep has more than doubled in size since 1971, and the original 400 immigrant chub have spawned 10,000 descendants, packed cheek by fin in the murky water. The Fish and Wildlife Service, as well as its state counterpart, the California Department of Fish and Game, warn Ridgecrest officials that if they harm the chub's habitat they will be subject to criminal charges under the endangered species act.

Damon Edwards, Ridgecrest's city administrator, is the man in the middle. "I'm getting letters from the water-quality people saying if I don't do something they're going to fine the city $1,000 a day," he wails, "and letters from the Fish and Wildlife people saying if I do anything I'll go to jail."

Taking mercy on Edwards, the water-quality board grants Ridgecrest time to seek a universally satisfactory solution. That proves elusive. C. Robert Feldmeth of Claremont McKenna College, an authority on desert fish who serves as a consultant to the Navy, proposes that artificial modification of Lark Seep might help it to sustain its chub with much less water. Yet the Fish and Wildlife Service insists that Ridgecrest must establish a backup chub habitat before tampering with Lark Seep, just in case Feldmeth's proposal doesn't work.

Feldmeth suggests the wildlife service's concern for the chub may be a bit excessive. He points out that the species has existed for only some 10,000 years, that it is inedible and not even particularly attractive. Raymond J. Bransfield of the service concedes that the chub "are not very pretty," but asserts that "they're pretty nifty for being able to survive where they do."

The plot awaits a resolution. Ridgecrest officials are pondering their next move, and puddles continue to plague China Lake. Is the Navy tempted to improvise a finale by "accidentally" dropping a bomb on Lark Seep? Beverly Kohfield, a Navy biologist, admits "there have been some jokes to that effect." Yet the chub has won the affection of at least a few China Lake employees; operators of a lunch wagon on the base often feed the fish their leftovers, and they have dubbed one remarkably, well, chubby fellow "Moby Chub." Can he act? —*J.H.*

MOHAVE TUI CHUB swarm in Lark Seep, a pond on the China Lake Naval Weapons Center in California. Lark Seep sustains the largest known population of the fish.

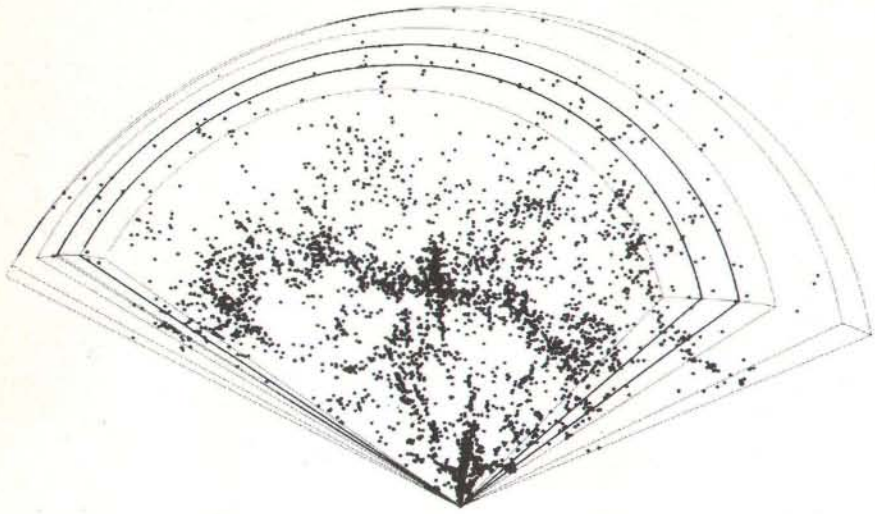
PHYSICAL SCIENCES

Up Against the Wall

Cosmologists take a few lumps from the latest cosmic maps

The end of 1989 offered a truly cosmic irony: just as the Berlin Wall was crumbling here on the earth, two astronomers announced their discovery that nature has built

Geller and Huchra claim to have discovered the largest structure known in the universe



***GALAXIES** in four recently surveyed "slices" of the universe are shown in a three-dimensional projection. Each slice is about 500-million light-years deep. The horizontal band of galaxies running across the map is the "Great Wall."*

an enormous wall of its own—in this case, a wall of galaxies that seems to be a staggering 500-million light-years across. They say it is the largest structure yet observed in the universe.

Astronomers Margaret J. Geller and John P. Huchra of the Harvard-Smithsonian Center for Astrophysics found this structure, the "Great Wall," while they were plotting the distribution of galaxies in space. It is the most recent of a series of discoveries that collectively seem to indicate that the universe is a lumpier and therefore more confusing place than cosmologists have suspected.

The first indication that matter is not uniformly distributed throughout the universe, even over very large scales, came in 1981, when a team of astronomers led by Robert P. Kirshner of Harvard University found a surprising 100-million light-years wide bubble, quickly dubbed the Hole in Space, that appears to contain less than one fifth the cosmic average density of galaxies. The Great Wall represents an opposite extreme: this structure contains roughly five times the average density of galaxies and, despite its great width, appears to be no more than 15-million light-years thick. The full extent of the wall is still unknown, because it runs off the edges of the regions mapped by Geller and Huchra.

Progressively larger surveys of galaxies have revealed progressively larger structures, and there is no end in sight; Geller and Huchra note that the "size of the largest structures we detect is limited only by the extent of the survey," and current surveys cover only about one hundred-thousandth of the volume of the visible universe. As they continue to work on their full-sky survey—which will ultimately plot some 15,000 galaxies—larger features may well emerge.

Such enormous structures pose a quandary for current models describing how matter became organized in the universe. They are too large to have formed as a result of gravitational clumping unless there were already significant lumps present in the universe at its birth. And yet the cosmic microwave background, which is believed to be the "echo" of the big bang that began the universe, appears to be extremely smooth. Anthony C. S. Readhead and his colleagues at the California Institute of Technology have found that the background is uniform to one part in 60,000 on the scale of two arc minutes—about the size that a galactic cluster would appear at the distance at which the background was emitted. The recently launched *Cosmic Background Explorer* will refine these limits considerably in the coming months [see "The Cosmic Background Explorer," by Samuel Gulkis, Philip M. Lubin, Stephan S. Meyer and Robert F. Silverberg; SCIENTIFIC AMERICAN, January].

Any major lumps in the primordial universe would presumably show up as fluctuations in the background, but each attempt to measure these fluctuations sets smaller limits on their possible sizes. How could enormous objects such as the Great Wall form in an initially homogeneous universe? Cosmologists' standard answer is to invoke "dark matter," unknown particles that cannot be seen but that are believed to make up more than 90 percent of the mass of the universe. The existence of dark matter is inferred from the motions of galaxies and from models of the big bang that suggest the universe is dense enough to halt its current expansion. Lumps in the dark matter would not show up in the microwave background, but they could have provided the gravity necessary to get galaxy formation rolling.

Yet there are limits even to the size of structures that can be explained by dark matter, limits that the Great Wall seems to be pushing. Models that posit slow-moving dark particles—"cold dark matter"—have difficulty producing such large structures, whereas models that assume fast-moving dark particles—"hot dark matter"—do not accurately mimic the smaller-scale details seen in the universe. Geller and Huchra say their survey "poses a serious challenge for all current models."

Most cosmologists are struggling to modify their theories to fit with the latest observations. Others, such as Jeremiah P. Ostriker of Princeton University, are developing alternative models that do away with the enigmatic dark matter. All agree, at the very least, that current theories are far from complete. In the heavens as on the earth, it has not been a good season for orthodoxy. —*Corey S. Powell*

Roving Stones

A landmass was wandering over three billion years ago

Rock remembers its travels. During the formation of a volcanic or a sedimentary rock, trapped crystals of iron oxide record the direction of magnetic north. If plate-tectonic motions later carry the landmass closer to or farther from the Equator, the tilt of the magnetism recorded in successively younger rocks will change; the pole will appear to have wandered. In the 1950's evidence of apparent polar wandering provided some of the first hints that continents might not be fixed. Now paleomagnetism—the study of rocks' ancient magnetism—is providing direct evidence of plate motions during the earliest part of the earth's history.

Powerful dating techniques are extending paleomagnetism's reach into the Archean era, more than 2.5 billion years ago, when some workers think the large-scale horizontal movements of plate tectonics had not yet begun. Paleomagnetic results presented at the December meeting of the American Geophysical Union in San Francisco suggest that at least one landmass was in motion more than three billion years ago.

"We wanted to get paleomagnetic results from some of the oldest rocks in the world," says Paul W. Layer of the Geophysical Institute, University of Alaska at Fairbanks, and so he and his colleagues went to Swaziland, in southern Africa, where ancient crust is preserved in a geologic province called the Kaapvaal craton. For paleomagnetism to say anything about the history of a landmass, the studied rock must have remained locally fixed since its formation. Massive rock bodies are best, and so the workers chose three plutons—vast granite bodies formed when magma rose into the crust and cooled underground.

Layer and his colleagues drilled hundreds of core samples from the plutons, which have been exposed by erosion. With Michael McWilliams of Stanford University, Layer recorded the samples' magnetic vectors during heating to above 500 degrees Celsius, the temperature at which the rock's original magnetism would have been locked in as the newly emplaced pluton cooled. The result was a paleomagnetic pole for each pluton: a snapshot of its original orientation in the earth's magnetic field. To assemble these snapshots into a paleomagnetic history, however, the workers needed a precise date for each one.

Techniques based on the decay rate of uranium to lead or of radioactive potassium to argon can now date an Archean sample with an uncertainty of less than 10 million years. Carefully modulated heating, often by laser, releases the elements from a mineral sample; a sensitive mass spectrometer then detects the ratio of parent to daughter atoms, from which the sample age can be calculated. The analysis can be done on a single crystal grain of mineral, which ensures that the result is not simply an average of grains with varying, possibly aberrant ages.

The decay clock for a newly crystallized mineral starts ticking when it cools to a "blocking" temperature, at which the decay product becomes trapped in the crystal lattice. Granite, however, contains a potpourri of minerals—zircon, hornblende, biotite and feldspar—whose blocking temperatures range from 700 to as little as 200 degrees C, and the granite in a pluton may cool very slowly, in some cases taking tens of millions of years. With different minerals giving very different dates, how does one date the magnetism of the rock as a whole?

Layer and his colleagues took advantage of the disparate dates to reconstruct a complete history of each pluton's cooling. Uranium-lead dates on zircon crystals, determined by Alfred Kröner of the Johannes Gutenburg University in Mainz, West Germany, dated the cooling curve's starting point at 700 degrees C; argon dates for the other minerals, determined in Derek York's laboratory at the University of Toronto by Layer and Margarita Lopez-Martinez, filled in the cooling histories. Dating a pluton's magnetism then became a matter of tracing the cooling curve down to about 500 degrees C and finding the corresponding point on the time scale.

The workers assembled their dated poles, together with poles from less precisely dated neighboring formations, into a history of the Kaapvaal craton's changing magnetic orientation between about 3.2 billion and 2.7 billion years ago—a crude motion picture, they believe, of the craton's travels. The landmass seems to have wandered north and south by an average of about 17 millimeters a year—about as fast as plates move today. Layer and his colleagues are now refining that estimate through more detailed studies in southern Africa.

How long had continents or their ancient nuclei been drifting about the planet? York and Chris M. Hall of Toronto are contemplating similar studies of the oldest known rocks on the planet, in the Slave province of northwestern Canada. Uranium-lead analysis by other workers last year yielded an age for the rocks of 3.96 billion years—only half a billion years younger than the earth itself. —*Tim Appenzeller*

BIOLOGICAL SCIENCES

Cold Storage

Winter-proof critters suggest ways to store human organs

Evolution has endowed many creatures with a variety of biochemical tricks that let them survive when their body temperatures fall below zero. Could these strategies help to preserve human organs?

Insects, the most impressive animal winterizers, generally make glycerol to avoid the effects of freezing. This syrupy, sweetish alcohol works in two ways. First, it lowers the freezing point of body fluids by increasing the number of dissolved molecules. The housefly, so protected, can withstand temperatures as low as −10 degrees Celsius. Second, glycerol helps to transform water into a vitreous solid, or glass, instead of ice crystals, which would rupture cells. This effect enables the Arctic willow gall to survive temperatures of −55 degrees C.

Fishes and some other vertebrates have also evolved chemical protectants against the cold. A frost-tolerant frog, for example, makes vast quantities of glucose to control the process by which it freezes, says Kenneth B. Storey of Carleton University in Ottawa. The glucose raises the concentration of aqueous solutions in tissues that would otherwise be dehydrated by the thickening of blood, as the blood's water is locked up as ice.

All of this evolutionary inventiveness has not been lost on medical researchers. One is Gregory M. Fahy, leader of the organ cryopreservation project of the American Red Cross in Rockville, Md. Until recently, he says, workers seeking to preserve tissue through extreme chilling had succeeded only by blasting tiny amounts with liquid helium or by throwing a few cells at a time onto chilled mirrors.

Fahy and his co-workers wondered whether they could vitrify an organ as large as a rabbit kidney at relatively slower speeds if they first perfused it, via its vascular system, with a cocktail of chemicals including propylene glycol, a close relative of glycerol. The investigators kept each of the agents at a concentration just low enough to be tolerated and just high enough to promote vitrification under a pressure of 1,000 atmospheres. They then were able to vitrify the kidney by immersing it in a bath of isopentane at −130 degrees C, even though this produced a chilling rate that would have been too slow—absent the pressure and chemical agents—to avoid crystallization. "We're right on the edge of what we can just tolerate in cryoprotectants and pressure," Fahy says.

To prevent the vitrified kidney from freezing as it returns to body temperature, the workers bathed the organ evenly with radio waves. A problem remains, however: chilling injury. "No one knows the mechanism, except that it's not crystal damage," Fahy comments. "It's a fundamental road-

block that we're trying to overcome."

The Red Cross has been working on the kidney storage project for 17 years because demand for the organs constantly exceeds the supply. Two related circumstances exacerbate the problem. Because the organs are perishable they must be used quickly; thus, at any one time the likelihood of finding a good donor-recipient match is small. Even the immunosuppressive drug cyclosporine has not been able to raise the half-life of transplanted kidneys much above seven years. "If we could preserve kidneys for a year and accumulate a bank of 6,000 organs, we could find optimal matches in more than half the cases," Fahy says. "That could extend the transplants' half-life to 12 to 14 years." —*Philip E. Ross*

Snakes in the Grass

Their escape behavior tells an evolutionary tale

Garter snakes exhibit a startling degree of variation in their protective markings. Some individuals sport bright stripes; others are spotted or dull. Most other animal species turn out, on inspection, to harbor similar genetic variation in traits thought to be adaptive. The observation poses a problem for evolutionary theorists: Why hasn't one optimal type crowded out all the others?

In *Nature*, Edmund D. Brodie III of the University of Chicago describes findings that support one possible resolution of the paradox: natural selection might favor several different *combinations* of genes, thereby preserving variability in individual genes. An animal's behavior, for example, might interact with its color or pattern in ways that favor several different combinations of marking and behavior. Brodie reports that selection seems to have acted in precisely that way for garter snakes.

Among snakes in general, spotted patterns are thought to serve as camouflage; they are most common in species that feign immobility when approached. Stripes, which make it hard to judge the speed of a moving object, are characteristic of species that flee when threatened. To determine whether such a correlation exists in garter snakes, Brodie worked with nearly 500 newly hatched garter snakes. Noting the surface markings, he then prodded them and recorded how fast and how far they slithered away, as well as how often they turned and froze—behavior that, in a camouflaged animal, can confuse a pursuer.

Behavior and pattern were clearly genetically linked: siblings resembled each other in both respects. Furthermore, striped individuals tended to flee, whereas spotted individuals tended to freeze.

By demonstrating a (presumably adaptive) association between behavior and pattern within natural populations of a single species and showing that it has a genetic basis, Brodie has illustrated one way that genetic variability in such populations might be maintained. He may also have made life more complicated for those biologists who, for the sake of simplicity, sometimes consider physical features and behavior as if they were unconnected. —*T.M.B.*

The markings of garter snakes vividly illustrate the genetic variation in a wild population

***GARTER SNAKES** have abundant and conspicuous genetic variation. Experiments suggest the variability is related to genetic variation in defensive behavior.*

Kissing Cousins

A DNA repair system stops species from interbreeding

Old jokes to the contrary, if you cross a gorilla with a duck, you don't get anything. Species cannot interbreed successfully because of differences in their DNA: the offspring of such unions die, or they are sterile because they cannot combine the dissimilar maternal and paternal DNA strands. This latter obstacle holds even for closely related species whose DNA sequences differ by as little as 10 percent. Now two groups working independently have identified a molecular mechanism that prevents incompatible DNA strands from combining. Their discovery sheds light on what keeps closely related species from interbreeding and on how new species can arise through evolution; it may also explain how cells discourage some cancerous changes in their chromosomes.

Preventing species from crossbreeding seems to be a job of the piece of molecular machinery called the mismatch repair system, which is already known to serve a related purpose in the normal replication of DNA. The mismatch repair system consists of a set of proteins that detects errors in the base sequences of freshly made DNA strands. If a new strand does not complement the template strand on which it is modeled, the system cuts out the mistake and a piece of the surrounding DNA. Polymerase enzymes then make a replacement for

the excised DNA [see "The High Fidelity of DNA Duplication," by Miroslav Radman and Robert Wagner; SCIENTIFIC AMERICAN, August, 1988].

Miroslav Radman and Christiane Rayssiguier of the Jacques Monod Institute in Paris and David S. Thaler of the University of Utah had hypothesized that the mismatch repair system might also be involved in inhibiting recombination between the divergent DNA sequences of different species. To test their theory, they attempted to cross two species of bacteria. (Bacteria engage in a form of sexual reproduction called conjugation in which they exchange genetic information.)

The bacteria *Escherichia coli* and *Salmonella typhimurium* diverged in evolution about 150 million years ago, and today their DNA sequences differ by about 20 percent. Consequently, these species cannot conjugate successfully under normal circumstances. Working with mutants of these bacteria in which the mismatch repair system was defective, however, Radman and his colleagues were able to increase the rate of recombination up to 1,000 times, according to their report in *Nature* last November.

In evolution, Thaler explains, the mismatch repair system may help create new species by erecting a reproductive barrier between groups of organisms with slightly different DNA even when there are no geographic obstacles to their interbreeding. Random mutations would gradually make distinct populations drift apart genetically; at some point the genetic differences would become significant enough for the mismatch repair system to abort a critical number of recombination events and prevent cross-fertilization. The populations would then become separate species.

Ping Shen and Henry V. Huang of the Washington University School of Medicine also have demonstrated the role of the mismatch repair system in regulating recombination. They worked with slightly different copies of genes for immunoglobulin proteins taken from a mouse. When copies of these genes were inserted into normal *E. coli*, no recombinations took place. In *E. coli* with defective mismatch repair systems, however, the immunoglobulin genes recombined frequently. Commenting on Shen and Huang's work, Thaler suggests that the mismatch repair system may help prevent undesirable recombinations between repeated gene sequences in chromosomes; this activity may prevent chromosomal rearrangements that can lead to cancer. —*John Rennie*

MEDICINE

When Dad Drinks

Can his liquor intake impair his future offspring?

Signs in bars warn of the harm that pregnant women can inflict on their fetuses by drinking alcohol. Might a male also impair the cognitive ability of his future offspring by even a short period of heavy drinking before conception? Experiments performed with rats at the Washington University School of Medicine suggest such a scenario.

Theodore J. Cicero and David F. Wozniak reported their results at a recent neuroscience meeting in Phoenix. For 39 days the researchers fed 15 male rats a liquid diet consisting of 6 percent alcohol, enough to keep them continually intoxicated. The rats were then weaned from the alcohol diet and two weeks later were mated with a group of teetotaling females.

Male offspring of these unions opened their eyes and gained weight at the same rate as a group of control pups; they also performed most physical and perceptual tasks with the same proficiency. Yet the experimental offspring showed a marked deficiency in their ability to navigate through a maze to a food reward; they typically took about 50 percent more time to master the task than the control rats did.

Cicero and Wozniak tested only male pups because previous work by Cicero had indicated that sons of alcohol-consuming male rats had hormonal imbalances that did not appear in their sisters. "The animals look totally normal," Cicero says, "but then you look at specific areas [such as learning and hormonal levels] and you see abnormalities."

The rats serving as fathers in the experiments were young—the equivalent of adolescent human males. Cicero says he used young rats because his earlier studies had indicated that they are far more sensitive to the effects of alcohol than their elders. In future studies Cicero and Wozniak plan to use mature rats as fathers and to test the cognitive abilities of both female and male offspring.

The results obtained so far correspond to findings concerning the sons of human alcoholics, according to Cicero. Researchers have reported, for example, that sons of alcoholics are more likely to have hormonal abnormalities and to perform less well in school than either their sisters or the children of nonalcoholics.

These human studies are fraught with unknown genetic and environmental variables and so are difficult to interpret. But Cicero and Wozniak think their animal experiments indicate that alcohol can exert a mutagenic effect on the sperm of a prospective father and thereby affect his future offspring. The workers acknowledge that they do not know how the mutation occurs or how it causes either hormonal imbalances or cognitive deficiencies in male pups. "I'd love to have figured out the mechanism before going public," Cicero says, "but I think it's important to get this observation out there."

Jack Mendelson of Harvard Medical School agrees, calling the finding extremely important. "That some changes affecting cognitive development are transmittable through the male will stimulate huge interest," he remarks. —*J.H.*

Playing the Numbers

Do bodily rhythms play a key role in cancer survival?

Premenopausal women who undergo breast cancer surgery around the middle of the menstrual cycle are only one fourth as likely to relapse and die as those whose surgery is performed within a week before or after menstruation. This claim could reassure or terrify patients and demolish physicians' schedules. Is it true?

Late last year William J. M. Hrushesky of the Albany VA Medical Center in New York and his colleagues announced such a finding. They came under immediate fire from other cancer researchers who claim that their data show no such correlation. Regardless of who is right, the controversy may point up some of the limits of statistical studies in medicine.

The outline of Hrushesky's work is fairly simple: starting from evidence that the vigor of women's immune systems waxes and wanes over the course of the menstrual cycle and that immune responses may affect the spread of tumor cells, he and his colleagues embarked on five years of animal experiments. They found that female mice with breast cancer tumors survived the longest when the primary tumors were removed around the time of ovulation (corresponding to midcycle in women).

Meanwhile they also began a search

for human data that could support their hypothesis. The main problem they faced was finding breast cancer studies in which physicians had recorded the date of the last menstrual period before surgery. The correlation between timing and survival was obvious once the data were assembled, Hrushesky says: "If it were a subtle effect, we never would have found it in just 44 patients."

The effect was apparently too subtle, however, to show up in data from two other breast cancer studies whose investigators dispute Hrushesky's findings. In a British study of 81 patients, women who underwent surgery near menstruation appeared to survive longer than those operated on at midcycle (although the difference was not statistically significant). In addition, researchers involved with the international Ludwig breast cancer treatment trials reported data on 245 women that show no significant timing effect. (Hrushesky reported being unable to obtain menstrual data from the Ludwig trials; Richard D. Gelber of the Dana-Farber Cancer Institute in Boston, who was involved in the trials, says he does not remember Hrushesky's request.)

So who's right? Hrushesky says that the other studies might not show any correlation for any number of reasons: the patients in them apparently suffered from more advanced tumors, underwent more aggressive therapy and died sooner than those he and his colleagues followed.

In response, Gelber levels similar comments in the opposite direction. Hrushesky's animal experiments give plausibility to the idea of a menstrual link, he says, but as for the human data, "if you look at a lot of data sets [in enough ways], there's always a chance that something will come up." Statistics, Gelber says, measure the likelihood that a particular result is due to chance, but they do not reveal whether a real effect is operating. Out of even 20 studies of random data, he notes, at least one should yield results that have only a 5 percent probability of being due to chance. Subtle factors not in Hrushesky's model could be mimicking the effects of a menstrual link, Gelber says.

While the researchers argue over statistical methods, breast cancer patients will continue to undergo what is in effect an enormous randomized trial. Yet chronobiology (the study of how daily, monthly and other bodily cycles affect responses to disease and therapy) has already taken hold in other areas of treatment. During the past decade clinicians have established, for example, that the toxic effects of certain anticancer drugs can vary tenfold depending on whether the drugs are given in the morning or the evening. Tumor-killing power shows similar patterns. In one study 15 patients with ovarian cancer who underwent chemotherapy at random times of the day survived on average less than a year and a half, whereas 11 treated in a particular morning-evening pattern typically survived almost seven years.

Hrushesky and others contend, in fact, that chronobiology will be the next important advance in therapy for cancer and possibly for other diseases as well. In a utopian future, implantable pumps could deliver drugs according to optimal dosing schedules too complex for either doctors or patients to follow. At present, however, no one knows precisely how chronobiological rhythms affect the body's response to disease and to medications. —*Paul Wallich*

The Oncogene Connection

Proto-oncogenes encode proteins with a neural role

When the brain acquires a new thought or memory, biochemical processes go to work etching its trace into neurons. Although no one claims to understand what these processes are, neuroscientists assume that some of them must involve the activity of genes that direct the synthesis of new proteins—proteins that alter the structure of nerve cells or change the way they interact with one another. Researchers at the Roche Institute of Molecular Biology in Nutley, N.J., now think they have found some candidate genes. Tom Curran, one of the principal investigators, reports that two well-known proto-oncogenes (benign forms of genes that can cause cancer) play a key role in directing nerve cells to synthesize an important neurotransmitter in response to stimulation.

The connection between proto-oncogenes and neurons is not altogether surprising, Curran says. Proto-oncogenes in other types of cells are known to mediate the transcription of proteins from genetic instructions. The prospect that proto-oncogenes might be doing the same in neurons is drawing many neuroscientists into studying them. Curran described the work at the Society of Neurosciences' annual meeting in Phoenix last November.

Curran, an oncogene specialist, says he got into neuroscience "by the back door." He teamed up several years ago with James I. Morgan, a neurobiologist at Roche, to search for proto-oncogene activity in neurons. They initially studied epileptic seizures in mice, reasoning that if proto-oncogenes are activated by nerve stimulation, the effect should be most pronounced after a seizure. And indeed, the seizures induced a massive increase in messenger RNA for the proto-oncogenes c-*fos* and c-*jun* in the animals' hippocampal neurons. "The changes occurred within about 15 minutes of the seizure—much more rapidly than people had believed possible," Curran says. The effect is so striking that epilepsy researchers are now seeing whether probes for c-*fos* and c-*jun* can be employed to map the cells involved in seizures. The cells also showed a surge in the expression of proenkephalin mRNA, which encodes a precursor of the enkephalin neuropeptides, some of the brain's "natural" opiates.

Is there a direct link between c-*fos*, c-*jun* and proenkephalin? Curran and Morgan think there is. The protein products of the proto-oncogenes pair up and bind to a site on DNA that regulates the expression of the proenkephalin gene, which suggests that the increase in the proteins causes proenkephalin to be synthesized in greater quantity. The researchers propose that a seizure causes neurons to deplete stores of proenkephalin and simultaneously induces the expression of c-*fos* and c-*jun*, thereby stepping up synthesis of new proenkephalin.

Curran and Morgan think that the proto-oncogenes may have another function: mediating long-term changes in neurons, such as those that accompany learning or adaptation. The workers speculate that the expression of c-*fos* and c-*jun* represents an early stage in a long-term change in neuron structure; the proto-oncogene proteins may regulate the expression of other genes encoding structural and receptor proteins, as well as neurotransmitters such as enkephalin. Indeed, several groups have reported increases in c-*fos* and c-*jun* after neurons are stimulated with the neurotransmitter glutamate, which has been implicated in learning and memory.

Others think the experiments do not yet support such ideas. "It's possible that proto-oncogenes are routinely expressed during cell growth or recovery after a seizure or high-frequency stimulation," points out Daniel L. Alkon, a neuroscientist at the National Institutes of Health. "If that's so,

then proto-oncogene expression may be more characteristic of a recovering cell than of a learning cell." Yet Alkon agrees that work on proto-oncogenes is likely to make profound contributions to neuroscience. As he puts it, "It makes sense that there could be a common mechanism underlying the genesis of tumors and neuronal adaptation." —*June Kinoshita*

OVERVIEW

Indecent Burial

Obstacles to the disposal of nuclear waste proliferate

As calls for a resurgence of nuclear power in the U.S. multiply (a recent *New York Times* editorial entitled "Revive the Atom" was typical), so do political, legal and practical obstacles to the long-term disposal of nuclear waste. All the major programs for burying waste—from literally hot uranium fuel rods to boots with a few specks of americium on them—face protracted delays. A lack-of-progress report follows.

Plans for a permanent repository for high-level waste have undergone the most serious setbacks. High-level waste includes spent fuel rods from power plants and by-products from the manufacture of nuclear weapons. In 1982 Congress ordered the Department of Energy (DOE) to find a site for an underground repository and to open it by 1998. In 1987 Congress pushed the opening date back to 2003 and told the DOE to consider henceforth only one site—Yucca Mountain, a bone-dry ridge abutting an Air Force base in Nevada.

Recently Secretary of Energy James D. Watkins announced that the repository can open no sooner than 2010. Some observers consider even that date optimistic. Probably the biggest barrier to the repository is opposition by Nevadans, including virtually every politician in the state. The state has refused to grant the permits needed for the DOE to drill exploratory shafts into Yucca Mountain. The DOE has asked the Department of Justice to sue the state and thereby force it to issue the permits. Nevadans, some of whom have compared the DOE to the Kremlin (pre-Gorbachev), have vowed to take the fight to the U.S. Supreme Court, if necessary.

Another legal encumbrance stems from the DOE's attempts to find an "integrating contractor" to help it manage the project. Last year the department selected Bechtel, Inc. Then an unsuccessful bidder, TRW, sued, claiming the DOE had been biased toward Bechtel. As part of its case, TRW pointed out that one of the DOE officials awarding the contract was a former Bechtel employee. Last summer a federal court ruled that the contract should go to TRW. The DOE is appealing; Watkins has also intimated that the DOE might proceed without an integrating contractor. TRW, naturally, has threatened to sue again unless the contract is fulfilled.

When, or if, the political and legal issues are settled, the DOE must still prove that Yucca Mountain can contain its deadly contents for 10,000 years, the standard set by the Environmental Protection Agency. How is the inspection to proceed? Even that is in dispute. The DOE had proposed excavating exploratory shafts with drills and dynamite. Recently the Nuclear Waste Technical Review Board, a group of independent advisers appointed by the President, pointed out that the water needed to lubricate the drill and the fractures caused by the dynamite could themselves threaten the proposed repository's integrity.

When *that* issue is resolved, the DOE must investigate serious questions about the site's natural suitability. Yucca Mountain lies near seismically active faults and a volcano that erupted less than 10,000 years ago; there are signs that the region's water table has been historically unstable and could rise far enough above its present level to invade the repository. If Yucca Mountain proves unsuitable—and the DOE has emphasized recently that it might—the department reports back to Congress for further instructions; there is no backup site or contingency plan.

Delays in the Yucca Mountain project have also blocked progress toward a temporary storehouse where high-level waste could be held until a permanent repository is ready. In the mid-1980's the DOE recommended building such a storehouse, called a monitored retrievable storage facility, at a site in Tennessee. State officials there loudly objected, however, and in 1987 Congress rejected the DOE's recommendation; it also stipulated that the DOE could not even consider new sites for a temporary facility until a site for a permanent repository had been approved. Of course, this approval has been indefinitely postponed—and indeed may never be granted. The DOE has asked Congress to allow work on a temporary facility to proceed regardless of the status of the permanent site. If Congress agrees (which is in doubt), the department still must find a state willing to house the nation's high-level waste "temporarily."

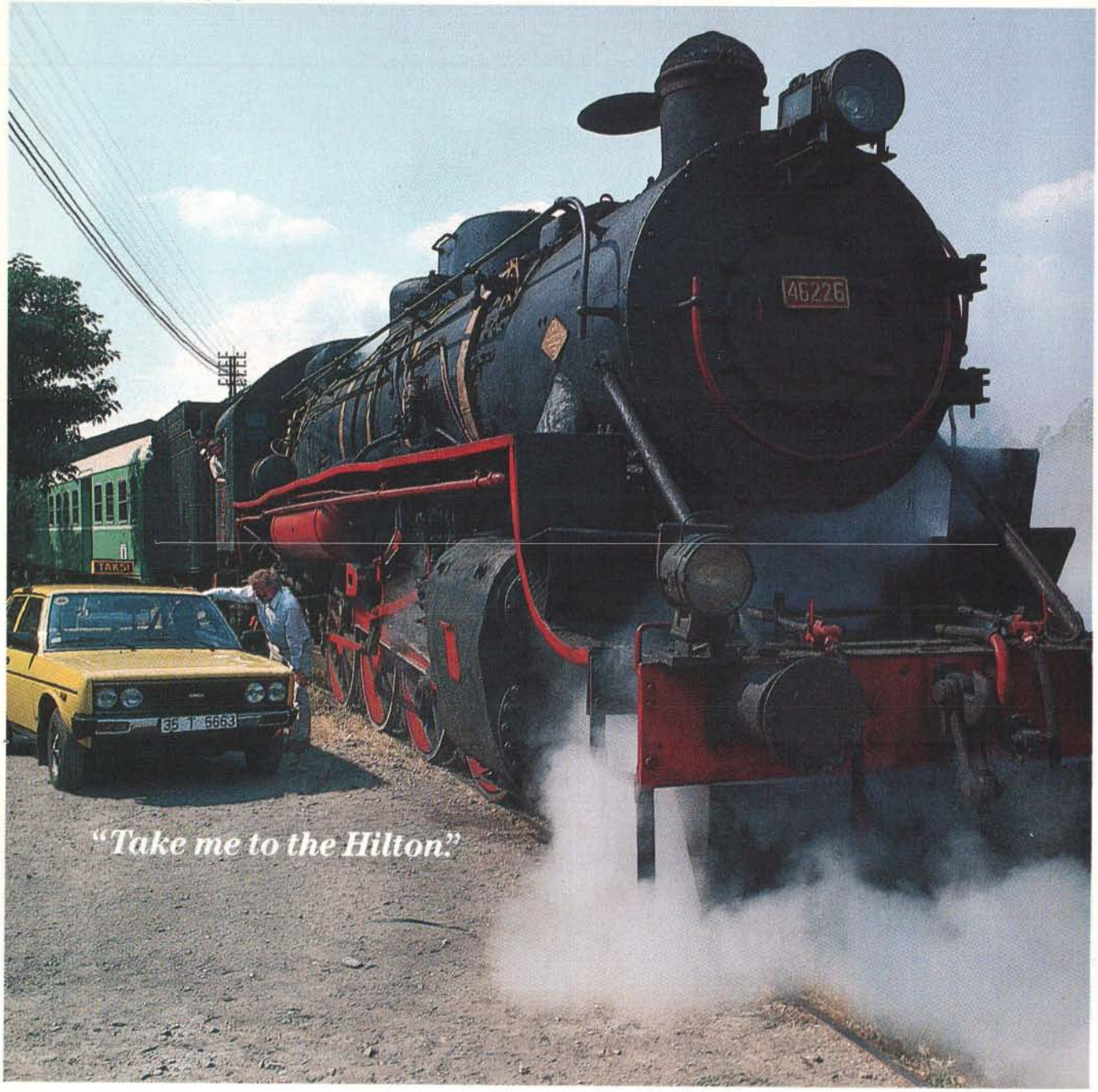
The U.S. has actually built one underground nuclear-waste repository. Called the Waste Isolation Pilot Plant (WIPP), it consists of a vast chamber carved into salt deposits near Carlsbad, N.M., and it is meant to hold so-called transuranic waste generated by DOE nuclear-weapons facilities. Transuranic waste contains radioactive elements heavier than uranium (such as plutonium); although the radioactivity emitted by the waste is less penetrating than that generated by high-level waste, it is long-lasting and still dangerous.

Begun in 1983, the WIPP was scheduled to start accepting deposits two years ago. But before the grand opening a panel of the National Academy of Sciences warned that salty water in the cavern might corrode waste containers and allow radioactive materials to leak into the surrounding aquifer. The panel recommended that the problems be studied further. After much debate the DOE recently agreed that the WIPP would accept only small amounts of waste for at least three years while tests proceed. Yet the department and its outside advisers have not been able to agree on exactly how much waste should be involved or on how the tests should be done. The experiments have yet to begin, and the WIPP remains empty.

Finally, there are the repositories for so-called low-level waste. Such waste includes everything from gloves worn by a nurse handling a short-lived radioactive tracer to reactor piping that will remain radioactive for hundreds of years. In 1980 Congress passed a law requiring states—either individually or jointly—to find permanent storage sites for low-level waste by 1986; in 1985, when it was clear that almost all the states would miss the 1986 deadline, legislators moved it back a decade. So far, only seven states have definitely found a long-term site: Alaska, Hawaii, Oregon, Utah, Idaho, Montana and Washington will all dump their waste in Hanford, Wash., which is already home to a DOE nuclear-weapons facility. At least one state, New York, has announced that it intends to challenge the constitutionality of the law on low-level waste, and several other states are reportedly considering similar action.

Radioactive-waste disposal, it would seem, remains an issue too hot to handle. —*John Horgan*

A Ladbroke Group Company



His explorations had taken him to some far-flung places, and it had been fun. Now he needed to collect his thoughts, his notes and several pieces of luggage. "Take me to the Hilton." He had a lot of time for the Hilton: appreciated its friendliness and efficiency. And shortly he would be appreciating one of the finer selections from their wine list. It was nice to know people you could be sure of. And good to be back.

◇ *For reservations at over 400 hotels, call your travel agent, any Hilton hotel or Hilton Reservations Worldwide. In the UK call (01) 780 1155.*

HILTON INTERNATIONAL

THE HILTON · THE HOTEL

SCIENTIFIC AMERICAN February 1990 Volume 262 Number 2

The Tragedy of Needless Pain

Contrary to popular belief, the author says, morphine taken solely to control pain is not addictive. Yet patients worldwide continue to be undertreated and to suffer unnecessary agony

by Ronald Melzack

"Pain," as Albert Schweitzer once said, "is a more terrible lord of mankind than even death itself." Prolonged pain destroys the quality of life. It can erode the will to live, at times driving people to suicide. The physical effects are equally profound. Severe, persistent pain can impair sleep and appetite, thereby producing fatigue and reducing the availability of nutrients to organs. It may thus impede recovery from illness or injury and, in weakened or elderly patients, may make the difference between life and death.

Sadly, there are some kinds of pain that existing treatments cannot ease. That care givers can do little in these cases is terribly distressing for everyone involved but is certainly understandable. What seems less understandable is that many people suffer not because their discomfort is untreatable but because physicians are often reluctant to prescribe morphine. Morphine is the safest, most effective analgesic (painkiller) known for constant, severe pain, but it is also addictive for some people. Consequently, it is typically meted out sparingly, if it is given at all.

Indeed, concern over addiction has led many nations in Europe and elsewhere to outlaw virtually any uses of morphine and related substances, including their medical applications. Even where morphine is a legal medical therapy, as it is in Great Britain and the U.S., many care givers, afraid of turning patients into addicts, deliver amounts that are too small or spaced too widely to control pain.

Yet the fact is that when patients take morphine to combat pain, it is rare to see addiction—which is characterized by a psychological craving for a substance and, when the substance is suddenly removed, by the development of withdrawal symptoms (for example, sweating, aches and nausea). Addiction seems to arise only in some fraction of morphine users who take the drug for its psychological effects, such as its ability to produce euphoria and relieve tension.

Furthermore, patients who take morphine for pain do not develop the rapid physical tolerance to the drug that is often a sign of addiction. Many people who are prone to addiction quickly require markedly escalating doses to achieve a desired change of mood, but patients who take the drug to control pain do not need sharply rising doses for relief. They may develop some tolerance initially, but their required dose usually rises gradually and then stabilizes.

I do not suggest that morphine be prescribed indiscriminately. I do urge lawmakers, law-enforcement agencies and health-care workers to distinguish between the addict who craves morphine for its mood-altering properties and the psychologically healthy patient who takes the drug only to relieve pain.

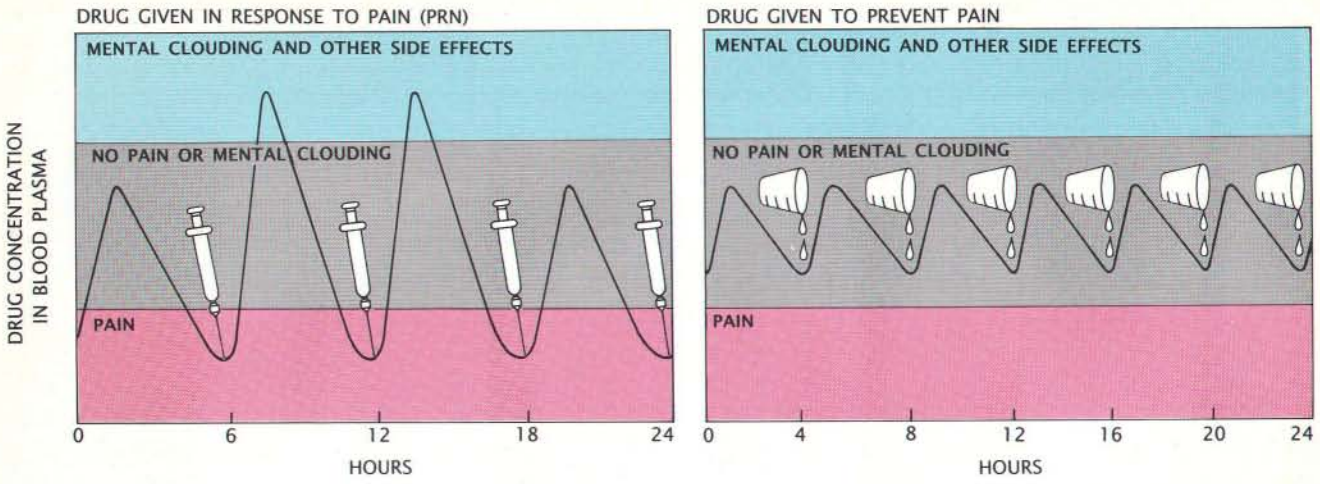
Morphine is a constituent of opium, which has been a medical therapy for longer than 2,000 years, since at least ancient Roman times. Opium is made by extracting a milky juice from the unripe capsule, or seedpod, of the poppy *Papaver somniferum* (grown abundantly in many Middle Eastern countries) and then drying the exudate to form a gum. This gum—the opium—can be eaten as is or added to a beverage.

By the 16th century opium was being carried by traders to Europe and the Orient. At about that time an opium-containing mixture called laudanum became a popular remedy in Europe for virtually all ailments. Later, smoking opium and tobacco together became yet another popular way to obtain the drug's benefits.

Soon after the turn of the 19th century, a young German pharmacist named Friedrich W. A. Sertürner isolated morphine from opium and identified it as opium's major active ingredient. Morphine's production was followed in 1832 by the isolation of yet another opiate, or opium derivative: codeine.

In the mid-19th century the introduction of the hypodermic needle made it possible to administer large amounts of drugs by injection. The

RONALD MELZACK, who has been studying the neurophysiology of pain for 35 years, is E. P. Taylor Professor of Psychology at McGill University and research director of the Pain Clinic at the Montreal General Hospital. After earning a Ph.D. in psychology from McGill in 1954 and accepting fellowships in the U.S. and abroad, he joined the faculty of the Massachusetts Institute of Technology in 1959. There, he and Patrick D. Wall began discussions that led to the 1965 publication of their now famous "gate control" theory of pain. He returned to McGill in 1963. This is his third article for *Scientific American*.



STANDARD APPROACH to morphine therapy for ongoing pain (*left*) calls for injections pro re nata (PRN), or "as needed." In practice this means injections are given only in response to pain; also, if the pain returns before four to six hours have passed, the patient often has to wait for help. By the time the next injection is delivered, the pain may be so severe that quite a large dose is needed, leading to mental clouding and other side effects, such as nausea. A more enlightened approach (*right*) seeks the actual prevention of pain and thus helps ease the fear of recurring agony. The morphine is given orally (in a dose tailored to the patient's needs) every four hours or even more frequently if a shorter schedule prevents pain more effectively. Because the doses are frequent, they typically can be relatively low, which reduces the incidence of side effects.

improved technology, which enabled a drug's effects to be felt quickly, led in many regions of the world to the ready prescription of injected morphine for severe pain. At the same time, more and more people began taking morphine for its emotional effects, and the number of addicts rose.

Eventually a search began for drugs that had morphine's analgesic properties but were not habit-forming. This quest resulted in the production of heroin, a synthetic compound similar in activity to morphine but soon found, disappointingly, to be quite as addictive. Various other opioids (chemicals with activity similar to that of opium) were then introduced, including methadone and meperidine (Demerol). Like the opiates, many of the opioids relieve pain, induce changes in mood and, unfortunately, are addictive to some extent.

Inevitably, the rising abuse of narcotics (by which I mean opiates and opioids) and of other mood-altering drugs spurred countries throughout the world to adopt antidrug regulations. At the same time, the extremely cautious administration of narcotics for pain became commonplace.

Today morphine therapy for pain is generally restricted to two groups of patients. It is prescribed over relatively short periods for hospitalized individuals who have discomfort caused by surgical incisions, and it is given over potentially longer periods to ameliorate the pain suffered by burn victims or people who have incurable cancer.

In many hospitals the standard prescription order says "PRN" (pro re nata, or "as needed"). This order essentially means that the drug is given only after pain returns. Typically, it is delivered by injection into a muscle or under the skin.

The result of the PRN approach is often a confrontation between the patient and the care giver, who expects morphine analgesia to last for four to six hours. The patient, whose pain has returned earlier than expected, is in agony and pleads to have the next injection. The health-care worker, fearful of causing addiction, refuses to comply. When the pain is finally treated, it may be so severe that a large dose has to be given, which increases the likelihood of side effects, such as mental clouding and nausea. Particularly when a patient has a terminal illness, the issue of addiction is meaningless, and delaying relief is cruel.

There is another, more humane way to treat pain, one that is slowly gaining acceptance. In this approach doses are given regularly, according to a schedule that has been actually tailored to prevent recurrence of the individual's pain. Thus, pain is controlled continuously; a patient does not wait for discomfort to return before receiving the next dose.

This enlightened, preventive approach evolved from pioneering work first undertaken some 20 years ago by Cicely M. Saunders, an English physician who established the first modern center devoted to caring for people who are dying of cancer or other diseases: St. Christopher's Hospice in London. Saunders urged physicians caring for terminally ill patients to face reality and palliate—to relieve pain, nausea and other discomforts—instead of making futile attempts to cure disease. The final days or weeks of a person's life, she believed, should be a time of peace and comfort, spent as pleasurably as possible in the company of family and friends.

To achieve this aim, Saunders prescribed the Brompton mixture, a version of a liquid analgesic that had been used for advanced cancer by several London hospitals, including the Brompton Chest Hospital, since the late 19th century. The mixture (made of morphine, cocaine, chloroform water, alcohol and flavoring syrup) had been eclipsed by injectable morphine, but Saunders realized that an orally delivered compound would allow many patients to spend a number of their last days at home; a visiting nurse would simply monitor them, making sure their pain was controlled.

Morphine has since been found to be the only important ingredient in the Brompton mixture, and so today patients who are treated with the preventive approach to pain take morphine alone, either as a tablet or mixed into a beverage. An initial dose of 10 milligrams is typically given and repeated every four hours. Then, over the course of perhaps several days or

weeks, the dose and timing are adjusted until a maintenance regimen is established that controls pain around the clock without producing mental clouding and other side effects.

For patients who have cancer, an approach emphasizing pain prevention is particularly wise. Pain and the fear of pain are perhaps their greatest source of suffering. In the early stages of the disease, some 50 percent of people have pain resulting from the cancer itself or from the procedures designed to arrest its spread. By the time the cancer has reached its final stages, about 70 percent of people report pain, which tends to be intense and persistent.

About 80 to 90 percent of cancer patients treated with the preventive approach obtain satisfactory relief, reporting that their discomfort is consistently bearable or, more frequently, gone. Roughly half of the remainder obtain relief with the addition of other therapies. This success rate is remarkable in view of the destructiveness of cancer and the severity of the pain associated with it.

Treatments continue to improve. There are now special capsules that release morphine slowly and so need to be taken only a few times a day. Also available are electronically controlled, portable pumps that deliver a steady infusion of medication under the skin.

Enough evidence has now been collected to demonstrate that the traditional, PRN approach, based as it is on the fear of addiction, makes little sense. Study after study of patients whose pain is most often treated with narcotics—namely, cancer patients, burn victims and those hospitalized for surgery—has shown that the patients who develop rapid and marked tolerance to, and dependence on, the narcotics are usually those who already have a history of psychological disturbance or substance abuse.

Let us first consider the problem of marked tolerance, which not only is a sign of possible addiction but is also a medical concern in its own right because the risk of side effects increases as the dose increases. For instance, delivery of extremely large amounts of morphine can induce coma and seriously impair respiration.

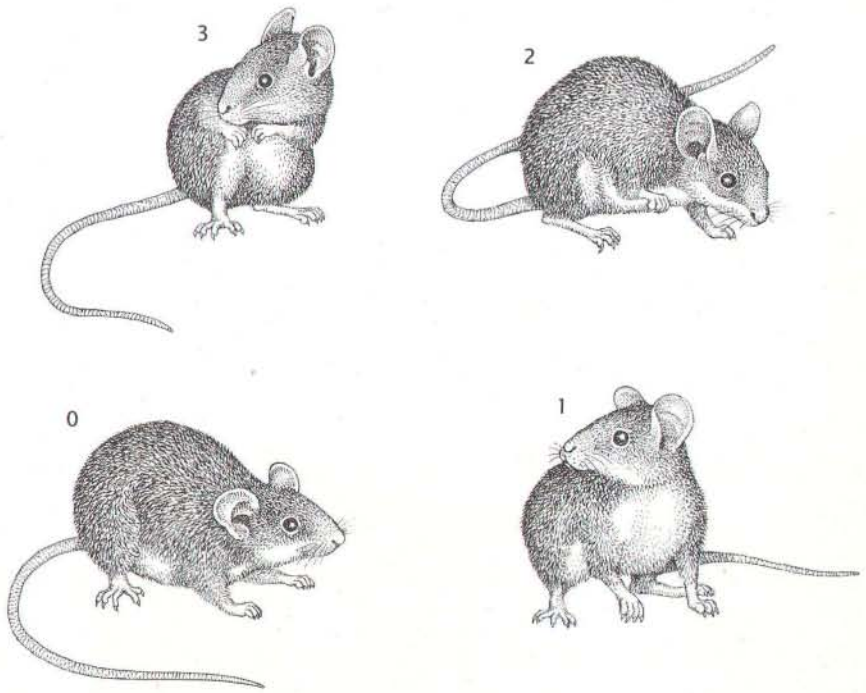
Robert G. Twycross, now at the Churchill Hospital in Oxford, England, has shown that relatively little tolerance develops in patients with cancer who take individually adjusted doses of heroin several times a day over long periods. The patients developed some tolerance to the drug initially, so that the doses had to be gradually raised over the first 12 weeks, but pain relief was achieved without producing serious side effects. Then the doses held fairly stable for months.

Balfour M. Mount, one of my colleagues at McGill University, and I recently found similar results when we studied tolerance to morphine in patients who spent more than a month in the Palliative Care Unit at the Royal Victoria Hospital in Montreal. (This unit, established by Mount, was the first service for palliative care at a large general hospital.) The patients in our study, who took the drug by mouth, answered a pain-evaluating questionnaire that I developed with Warren S. Torgerson of Johns Hopkins University. The overall intensity of the pain was ranked on a scale ranging from no pain (0) to pain that is mild (1), discomforting (2), distressing (3), horrible (4) or excruciating (5).

About 5 percent of the patients had persistently high pain levels (3 or higher). The remaining 95 percent had excellent pain control without requiring rapidly escalating amounts of morphine. Increase in pain, usually a sign of disease progression after a maintenance program has been established, was the most common reason for a rise in dose. Patients who found that their discomfort had decreased—either spontaneously or because of treatment, such as reduction of a tumor by radiation—usually required less medication.

John F. Scott of the Elizabeth Bruyere Health Center in Ottawa also uncovered little evidence of addiction when he analyzed many studies examining withdrawal symptoms in patients at cancer-treatment clinics. He reports that "if a cancer patient no longer requires a narcotic for pain control, a gradual reduction in dose will prevent any withdrawal symptoms, although these are usually mild or absent even after abrupt discontinuance." Any physical dependence is generally overcome without difficulty when doses are reduced over a period of days.

Studies of patients who received narcotics while they were hospitalized have also uncovered little evidence of addiction. In an extensive study



FORMALIN TEST measures the analgesic (painkilling) effects of medications on so-called tonic, or persistent, pain. A dilute solution of formaldehyde and saline is injected under the skin of a rat's paw, inducing pain that lasts for about 90 minutes. The rat licks its paw repeatedly, which is a sign of moderate pain (a pain rating of 3). Then, after a while, the animal holds the paw in the air (a rating of 2), steps on it gingerly (a rating of 1) and finally walks normally (a rating of 0). In this test, rats treated with morphine develop little tolerance to the drug's analgesic effect; that is, they do not require ever-increasing doses to obtain relief. This finding is consistent with the results of clinical studies showing that patients who take morphine for persistent pain do not acquire marked tolerance and do not become addicted.

Jane B. Porter and Hershel Jick of the Boston University Medical Center followed up on 11,882 patients who were given narcotics to relieve pain stemming from various medical problems; none of the subjects had a history of drug dependence. The team found that only four of the patients subsequently abused drugs, and in only one case was the abuse considered major.

Equally persuasive are the results of a survey of more than 10,000 burn victims. These individuals, who were studied by Samuel W. Perry of New York Hospital and George Heidrich of the University of Wisconsin at Madison, underwent debridement, an extremely painful procedure in which the dead tissue is removed from burned skin. Most of the patients received injections of narcotics for weeks or even months. Yet not a single case of later addiction could be attributed to the narcotics given for pain relief during the hospital stay. Although 22 patients abused drugs after they were discharged, all of them had a history of drug abuse.

Further evidence that narcotic drugs can be administered for pain without causing addiction comes from studies of "patient-controlled analgesia" in surgical patients and those hospitalized for burns. In such studies patients push a button on an electronically controlled pump at the bedside to give themselves small doses of morphine (through an intravenous tube). When these devices were introduced, there was considerable fear that patients would abuse the drug. Instead it soon became clear that patients maintain their doses at a reasonable level and decrease them when their pain diminishes.

Studies that explore how morphine produces analgesia are helping to explain why patients who take the drug solely to relieve pain are unlikely to develop rapid tolerance and become addicted. On the basis of such studies, my former student Frances V. Abbott and I proposed in 1981 that morphine probably has an effect on two distinct pain-signaling systems in the central nervous system and that one of these—which gives rise to the kind of pain typically treated with morphine—does not develop much tolerance to the drug.

Our proposal grew out of my efforts to develop a test in animals that would accurately determine the effectiveness of analgesic drugs on the kind of pain most often requiring narcotics in human patients: the prolonged, or "tonic," kind that persists long after an injury is suffered. This is the sort of pain that chronically bedevils cancer patients. When an injury first occurs, it gives rise to what is called phasic pain, which is brief and rapidly rises and falls in intensity. (The pain felt the instant a finger is cut would be called phasic.) Such phasic pain is usually followed by the tonic kind.

For many years investigators interested in measuring the analgesic effects of drugs subjected rats to what is called the tail-flick test. After a rat is injected with a test drug, its tail is immersed in hot water; the time between immersion and when the rat flicks its tail out of the water is measured as an index of pain. When morphine's effectiveness was examined with this test, investigators repeatedly found evidence of marked tolerance: the animals required ever-increasing doses in order to keep the tail in the water for a given time. Such results were interpreted to mean that human patients in pain would readily become tolerant to morphine and so would become addicted to the drug.

There is a major problem with the tail-flick test, however. It gives rise to suddenly rising, phasic pain, which is not the kind for which morphine is usually prescribed. To gain more information about the effects of analgesics on persistent, tonic pain in humans, John O'Keefe, David Dubuisson and Stephen G. Dennis, who were then my students, developed what is called the formalin test. A small amount of formalin—formaldehyde diluted in saline—is injected under the skin of a rat's forepaw. When the animal is not given an analgesic, the formalin produces moderate pain that lasts for about 90 minutes, as evinced by the animal's tendency to lick the paw and a reluctance to put weight on it. If a drug soothes the hurt, the animal puts weight on the paw more quickly.

With the formalin (tonic-pain) test, Abbott and I (later joined at McGill by our colleague Keith B. J. Franklin) discovered that rats developed relatively little tolerance to the analgesia produced by successive injections of morphine. The most logical explanation for the different degrees of tolerance found in the tail-flick and formalin tests was that phasic and tonic pain are invoked by two distinct neural systems that have differing tolerance to morphine.

Other lines of evidence added support to this idea. For instance, Dennis and I examined the effect on pain of several drugs that interact with morphine (or that alter pain in their own right) in both the tail-flick and the formalin tests. The results were striking. Drug effects that we found in one test were absent or even reversed in the other. For example, drugs that reduced morphine analgesia in one test either had no effect or enhanced the analgesic effect in the other test. If the neural systems that respond to phasic and tonic pain were one and the same, the effects of the drugs on morphine's activity should have been identical in both tests.

My colleagues and I think we now know which of the many neural pathways in the spinal cord and brain constitute the two pain-signaling systems that are sensitive to morphine. We also know something about their functioning and how they are affected by morphine. In both systems, information about pain is delivered to the dorsal horns (wing-shaped regions) of the spinal cord by peripheral neurons emanating from the skin and other body tissues [*see illustration on opposite page*]. Ascending neurons originating in the dorsal horns then relay the pain signals upward through the spinal cord to various parts of the brain.

The pain-signaling system that my colleagues and I think is most associated with sudden, phasic pain is called the lateral system. The name derives from the simple fact that the system's tracts, which project to the sensory cortex, pass through the brain stem at a position to the side of the brain stem's central core. The system that is probably responsible for persistent, tonic pain is called the medial system; its tracts pass through the central core of the brain stem.

Among the more salient properties of the lateral system are the rapid conduction of impulses and an organization that maps the relative position of body sites. These properties would enable the system to give rise to sudden, sharp pain in a readily identified spot in the body. Kenneth L. Casey of the University of Michigan at Ann Arbor and I have proposed that the lateral tracts also account to a great extent for the sensory qualities of pain, such as throbbing or burning.

The activity of the lateral system is apparently dampened rather quickly, which would explain why phasic pain often subsides promptly. The inhibition is accomplished by a system of neurons that originates in what is called the periaqueductal gray matter in the part of the brain stem known as the midbrain. This descending system sends signals downward to the dorsal horns, where it inhibits the transmis-

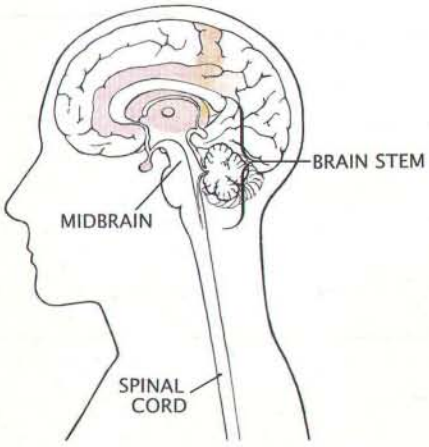
sion of pain signals from the peripheral nerves to ascending tracts. After an injury, it is apparently activated by the body's own opioids (endorphins and enkephalins).

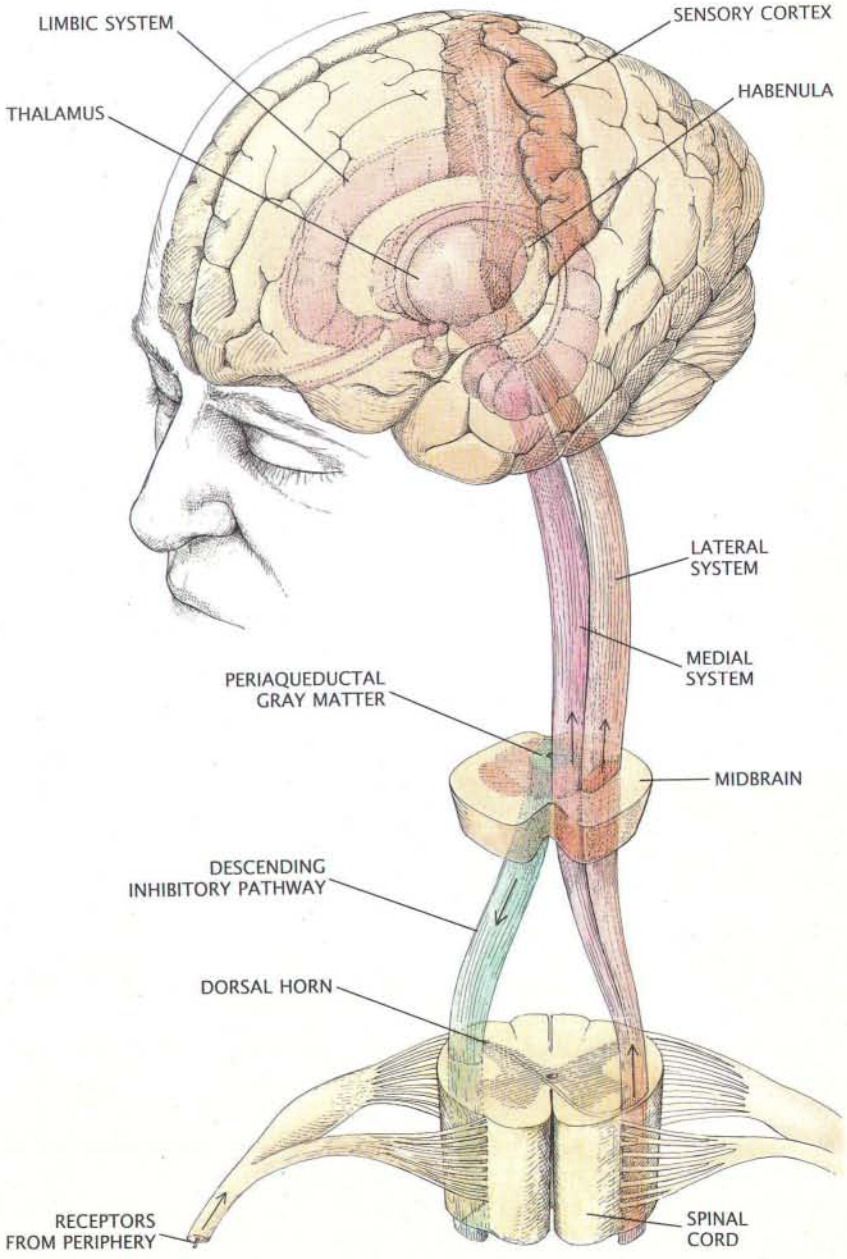
If, as we suggest, the lateral system carries the signals that give rise to sudden, phasic pain, then it is not surprising that the system is naturally subject to powerful inhibition. Sudden pain from a newly acquired injury could well overwhelm an animal, preventing it from fighting, running for cover or burrowing to escape a predator during an emergency.

The other pain-signaling system—the medial system—differs from the lateral system in many ways. For example, a number of its tracts send impulses to the limbic system, which comprises the subcortical regions of the brain involved in motivation and affect. Hence, we think the medial system controls the emotional component of pain, producing qualities one might describe as wretched, terrifying, vicious and the like. The system also influences the actions one takes in response to such feelings.

Because the medial system conducts signals relatively slowly through many small neurons, it is not well suited for providing precise information during emergencies. Instead it is more suited for producing diffuse, unpleasant feelings for some time after an injury has occurred. Such feelings would help ensure that, having survived an immediate threat, a wounded individual would feel miserable and so remain inactive long enough to heal.

Where does morphine exert its effects? In both the lateral (phasic-pain) and the medial (tonic-pain) systems, morphine clearly has some direct effect at the dorsal horns. It is also well known that morphine can activate the descending inhibitory system originating in the periaqueductal gray matter. Abbott and others in my laboratory have found that this descending





TWO SYSTEMS of neurons evoke pain: a medial system (*pink*), which passes through the central core of the brain stem, and a lateral system (*orange*). Both are bilateral, consist of several tracts and relay to higher centers the pain signals that come into the dorsal horns of the spinal cord. The medial system is thought to be most responsible for persistent (tonic) pain. Because it sends signals to the limbic system of the brain, which influences emotions, it is also believed to give rise to the affective component of pain (reflected by such descriptions as "frightful" or "cruel"). The lateral system is thought to be most active during phasic pain, which is sudden and sharp. Because it sends signals to the sensory cortex, it probably gives rise to such sensations as cramping or stinging. Morphine can inhibit both systems, but the medial (tonic-pain) system develops much less tolerance to the drug's analgesic effects than does the lateral (phasic-pain) system—which may explain why patients who take morphine for persistent (tonic) pain do not develop great tolerance to it. Morphine produces analgesia in part by inhibiting the flow of pain signals from the peripheral nerves to the ascending pathways; it acts directly at the dorsal horns and also activates a descending inhibitory system (*blue*) that originates in the midbrain. Morphine also acts at sites above the periaqueductal gray matter of the midbrain, including the limbic system and the habenula, which has strong links to both the medial and limbic systems. Such activity apparently contributes to the drug's analgesic effect on persistent pain.

system has a greater impact on the lateral system than on the medial system, which suggests that much of morphine's power over sudden, phasic pain is mediated by the descending neural tracts.

Morphine's analgesic activity certainly is not confined to the dorsal horns and the midbrain. For instance, strong evidence indicates that morphine acts on the limbic system, which is known to play a major role in both pain and pleasure. Such activity could well dampen the pain sensations produced by the medial (tonic-pain) system, which sends a great many impulses to the limbic system.

A recent study by S. Robin Cohen, my student, and myself lends additional support to the idea that morphine's influence over the medial system derives in part from activity above the midbrain. We injected morphine into the habenula, a small region of the brain (just behind the thalamus) that has strong links with the limbic system and a part of the medial system in the midbrain. The injections produced analgesia in the formalin test but not in the foot-flick test (similar to the tail-flick test), which suggests that morphine acts at the habenula and that, when it does, it inhibits the medial but not the lateral system. This finding and others indicate that more research should be devoted to areas above the midbrain if investigators are to gain a fuller understanding of how morphine eases persistent, tonic pain without inducing tolerance to repeated doses of the drug.

In view of the complexity of the neural mechanisms of pain, it is not surprising that morphine's ability to produce analgesia has been found to vary greatly from person to person. An important message emerging from studies of such variation is that the need for a high dose is not necessarily a sign of addiction.

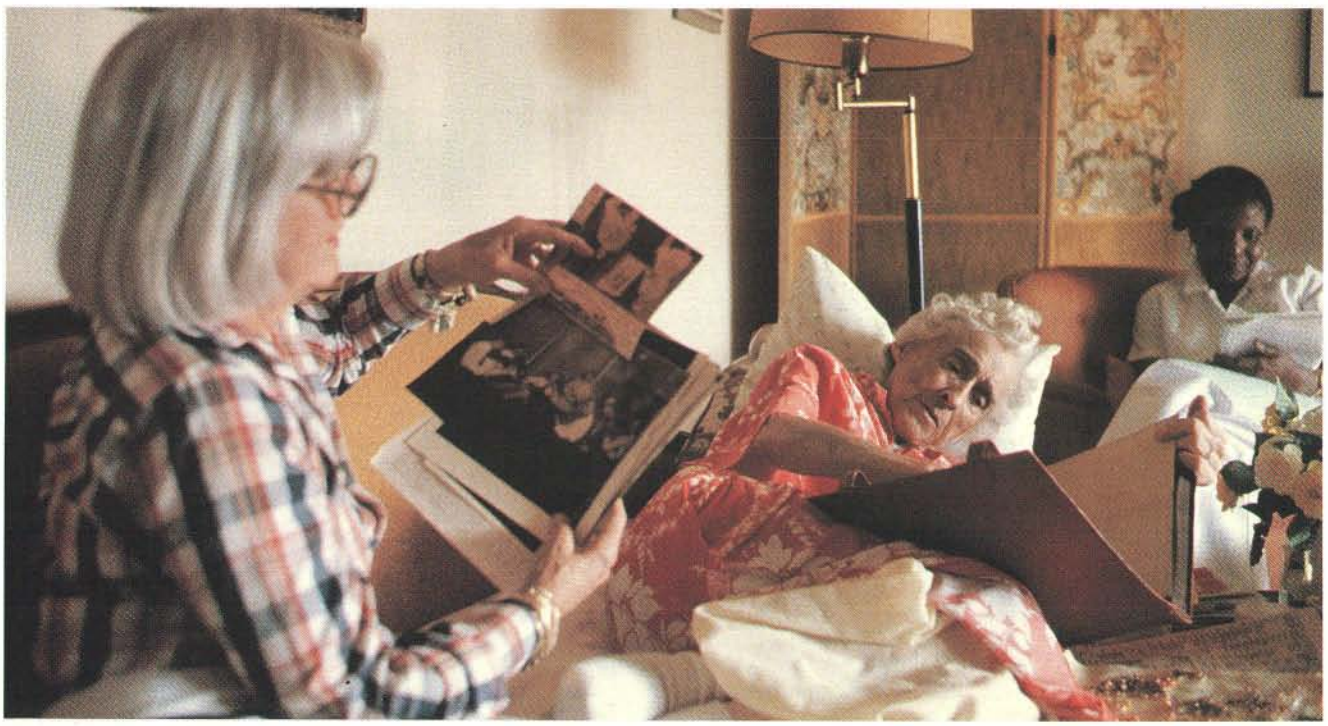
In one such study involving cancer patients, Robert Kaiko, now at the Purdue Frederick Company in Norwalk, Conn., and his colleagues at the Memorial Sloan-Kettering Cancer Center found that to achieve a given level of analgesia, less morphine was needed by older patients than by younger patients, and less was needed by blacks than by whites. Similarly, patients with dull pain needed less morphine than did those with sharp pain, and patients with stomach pain needed less morphine than did patients with pain in the chest or arm.

Genetic factors might also influence an individual's response to the analgesic power of narcotics, as Anthony L. Vaccarino (my student), R. Andrew R. Tasker, now at the University of Prince Edward Island, and I learned recently when we examined the effects of morphine and its antagonist naloxone in a strain of mice specially bred for studies of immunologic function. We unexpectedly found that the "antagonist" actually enhanced morphine analgesia and produced analgesia on its own in rats subjected to the formalin test. These surprising findings, which so far have been documented only for this strain of mice, are clearly the result of a genetic anomaly.

The discovery of a genetic influence on morphine's actions raises the possibility that susceptibility to addiction might also have a genetic component in some people. Evidence collected by other groups is consistent with that idea, although little work addresses the problem directly.

There is no way to identify patients who might be genetically predisposed to morphine addiction, but I must emphasize again that a person's psychological history is indicative of risk. More than 50 percent of narcotics abusers have had bouts of major depression, and 87 percent have a history of psychiatric disorder.

Society's failure to distinguish between the emotionally impaired addict and the psychologically healthy pain sufferer has affected every segment of the population. Per-



CANCER PATIENT (*reclining*), shown with a nurse (*left*) and a home health aide (*right*), is under the care of a hospice program. Such programs aim to palliate—to ease the pain, nausea and other discomforts of people who are terminally ill. Most hospices take the preventive approach to pain, which often enables patients to spend many of their final days at home.

haps the most distressing example is unnecessary pain in children.

Many health-care workers undertreat pain in youngsters, not only because of fear of addiction but also because of the mistaken belief that young children do not feel pain as intensely as adults. In a classic study, Joann M. Eland and Jane E. Anderson of the University of Iowa found in 1977 that more than half of the children from four to eight years old who underwent major surgery—including limb amputation, excision of a cancerous neck mass and heart repair—were given no medication for relief of their postoperative pain; the remainder received inadequate doses. When 18 of the children were matched with adults who underwent similar procedures, the children as a group were found to have been given a total of 24 doses of analgesic drugs, whereas the adults were given a total of 671 doses.

The elderly also pay the penalty of ignorance. In a study of postsurgical pain my colleagues and I found that surgical wards contain two basic populations: a young and middle-aged group that recovers quickly and an older group whose pain remains severe and lingers for many days beyond the normal three- to four-day recovery period. Despite the persistent, high level of pain in these older patients (presumably because of complications that arise after surgery) and despite the longer recovery period, they do not receive larger doses or a higher daily amount of medication. About 30 percent of the patients on a surgical ward at any time fall into this older category; they thus represent a substantial number of people who suffer needlessly high levels of pain.

The pain suffered by burn victims is known to be agonizing, and yet it, too, tends to be poorly controlled. Manon Choinière of the Burn Center at the Hôtel Dieu in Montreal and I found that even in the best burn facilities—those with highly capable, compassionate physicians, nurses, physiotherapists and others—pain levels are high. Our study of 30 consecutive patients who underwent debridement and physiotherapy (exercise to prevent loss of joint flexibility) classified the severity of pain on the basis of the pain questionnaire I developed with Torgerson. We discovered that during treatment in the first two weeks, 23 percent had severe ("horrible") pain, and 30 percent had extremely severe ("excruciating") pain. Even when the patients were at rest, 13 percent of them reported having severe pain, and another 20 percent said they had extremely severe pain. These data, by the way, were obtained from patients who were already medicated according to standard textbook recommendations (that is, the drug order said "PRN").

For many patients who are hospitalized for surgery or burns or who have terminal cancer, the prescription is clear: a preventive approach to pain should be instituted to maximize the effectiveness of narcotics therapy. What, though, should be done for people who suffer from debilitating chronic pain but who do not have a fatal illness? These people have traditionally been excluded from long-term therapy with narcotics, again for fear they would become addicts.

Consider the case of a 26-year-old athlete who sustained a major spinal injury that caused him to suffer from excruciating pain in the back and legs. The pain rendered him unable to work, and he became a burden to himself, his family and society, which pays his medical bills. His physician discovered that small doses of morphine taken orally each day (the way cancer patients receive them) obliterated the pain. With the help of the medication, the young man resumed working and made plans to marry his childhood sweetheart, who was accepting of his injury.

One day, however, the physician was accused by his regional medical association of prescribing narcotics for a purpose unapproved by the association and of turning the patient into an addict. Fearful of losing his medical license, the physician stopped prescribing the drug. (Where morphine administration is allowed by law, physicians can technically prescribe it at will, but they are in fact restricted by the regulations of medical societies, which control licensing.)

Of course, the young man's pain returned. In desperation, he turned to other physicians and was rebuffed. He then sank rapidly into depression and again became mired in helplessness and hopelessness.

It was once unthinkable to give narcotics indefinitely to patients who were not terminally ill. Yet studies designed to examine addiction specifically in such patients are beginning to show that for them, as for the standard candidates for narcotics therapy, these drugs can be helpful without producing addiction.

In one recent study Russell K. Portenoy and Kathleen M. Foley of Sloan-Kettering maintained 38 patients on narcotics for severe, chronic noncancer pain; half of the patients received opioids for four or more years, and six of these were treated for more than seven years. About 60 percent of the 38 patients reported that their pain was eliminated or at least reduced to a tolerable level. The therapy became problematic in only two patients, both of whom had a history of drug abuse.

With cautious optimism, Portenoy and Foley suggest that morphine might be a reasonable treatment for chronic pain in many patients who are not terminally ill. They point out the problems that may accompany narcotics maintenance therapy, and they provide careful guidelines for monitoring patients. Studies such as theirs are doing something in medicine that is akin in aeronautics to breaking the sound barrier. They represent a breakthrough to a reasoned, unbiased examination of the effectiveness of narcotics in patients who have rarely been considered for such therapy.

Among the critics of long-term narcotics therapy for such patients are physicians and others who fear that people will simply be given a prescription for a drug and will never receive the advantages of a multidisciplinary approach to the care of pain. Yet both approaches are compatible; in fact, they complement each other.

For the future, many more well-controlled studies are needed to provide data on the long-term effects of narcotics on chronic noncancer pain. At the same time, medical and government agencies must provide the authorization and funds for such studies to take place. The goal is nothing short of rescuing people whose lives are now being ruined by pain.

FURTHER READING

NARCOTIC ANALGESICS IN CLINICAL PRACTICE. R. G. Twycross in *Advances in Pain Research and Therapy*, Vol. 5. Edited by John J. Bonica et al. Raven Press, 1983.

CHRONIC USE OF OPIOID ANALGESICS IN NON-MALIGNANT PAIN: REPORT OF 38 CASES. R. K. Portenoy and K. M. Foley in *Pain*, Vol. 25, pages 171–186; 1986.

THE CHALLENGE OF PAIN. Revised edition. Ronald Melzack and Patrick Wall. Penguin USA, 1989.

TEXTBOOK OF PAIN. Second edition. Edited by Patrick D. Wall and Ronald Melzack. Churchill Livingstone, Inc., 1989.

INFLUENCE OF THE PAIN AND SYMPTOM CONTROL TEAM (PSCT) ON THE PATTERNS OF TREATMENT OF PAIN AND OTHER SYMPTOMS IN A CANCER CENTER. Eduardo Bruera, Carleen Brenneis, Mary Michaud and R. Neil MacDonald in *Journal of Pain and Symptom Management*, Vol. 4, No. 3, pages 112–116; September, 1989.

The Variable Sun

Its steady warmth and brightness are illusory; the sun's output of radiation and particles varies. Systematic observations are beginning to unveil the causes of these changes and their effects on the earth

by Peter V. Foukal

To someone lying on the beach or taking a daytime stroll, the blazing sun seems constant and unchanging. Actually, the sun is a variable star. The well-known 11-year "sunspot cycle"—which is thought to be now near its peak—is but one aspect of a complex, 22-year magnetic fluctuation during which the sun varies in its output of visible and ultraviolet light, X rays and charged particles. These fluctuations can heat and expand the earth's upper atmosphere, cause auroras, knock out power lines, alter the planet's ozone layer and perhaps influence climate. Even this cyclic variation cannot be considered reliable, because the sun has exhibited quite different patterns of behavior as recently as the 17th century, and there is good reason to expect that its behavior will change again.

This possibility is of more than academic interest, because any major change in the sun's luminosity or even in its level of activity could affect the habitability of the earth. Current discussions of possible changes in the global environment tend to focus on the effects of human activities, such as the climatic implications of accumulating greenhouse gases and the destruction of ozone by chlorofluorocarbons. Understanding and quantifying these effects demands an awareness of other causes of environmental change, particularly long-term variations in the sun's emission of light and charged particles. Investigators are working to determine more accurately the connections between conditions on the sun and those on the earth and to predict—or determine if it is even possible to predict—the future course of solar variability.

PETER V. FOUKAL is a solar physicist and president of Cambridge Research and Instrumentation, Inc., in Massachusetts. The author's ongoing research in solar physics includes infrared observations of the sun at Kitt Peak Observatory, development of an instrument that measures solar-plasma electric fields at Sacramento Peak Observatory and work on the sun's luminosity variation. Foukal received his bachelor's degree from McGill University in Montreal and his doctorate from the University of Manchester in England. Between 1969 and 1979 he was a research fellow and lecturer at the California Institute of Technology and then at Harvard University. His textbook, *Solar Astrophysics,* will be published this month by Wiley Interscience.

The first indication that such predictions might be possible appeared in 1843, when Heinrich S. Schwabe, a German apothecary and amateur solar observer, announced that the number of dark spots visible on the solar disk seemed to vary in a regular, roughly 10-year cycle. Schwabe's evidence for a sunspot cycle came to the notice of J. Rudolf Wolf, who became the director of the newly established Zurich Observatory in 1855. At Zurich, Wolf tracked the daily sunspot number based on reports from an international network of observers; he also compiled a history of the sunspot number based on records from the previous 150 years. Wolf found a corrected average period of about 11.1 years for the sunspot cycle, although both the period and amplitude varied considerably from cycle to cycle.

A graph of the sunspot cycle from 1610 to the present reveals that the sunspot number has oscillated without interruption since about 1715 [*see illustration on page 28*]. During the 13 cycles of reliable data collected since 1848, the cycle's length has varied between about 10 and 12 years. The cycle's amplitude has been even less regular, ranging from an annual mean spot number of about 45 in 1804 and 1818 to a peak of roughly 190 in 1957. The present cycle may produce the highest sunspot number and overall level of activity yet recorded, judging from its current behavior.

The regular variations of the sunspot cycle are notably absent for the years from 1645 to 1715, when very few sunspots were observed. This period of depressed solar activity is known as the Maunder minimum, named after the British solar astronomer E. Walter Maunder, who did much to draw attention to it in the late 19th and early 20th centuries. Astronomers initially tended to ignore Maunder's findings or to attribute them to the crude telescopes and poor observational techniques of the age.

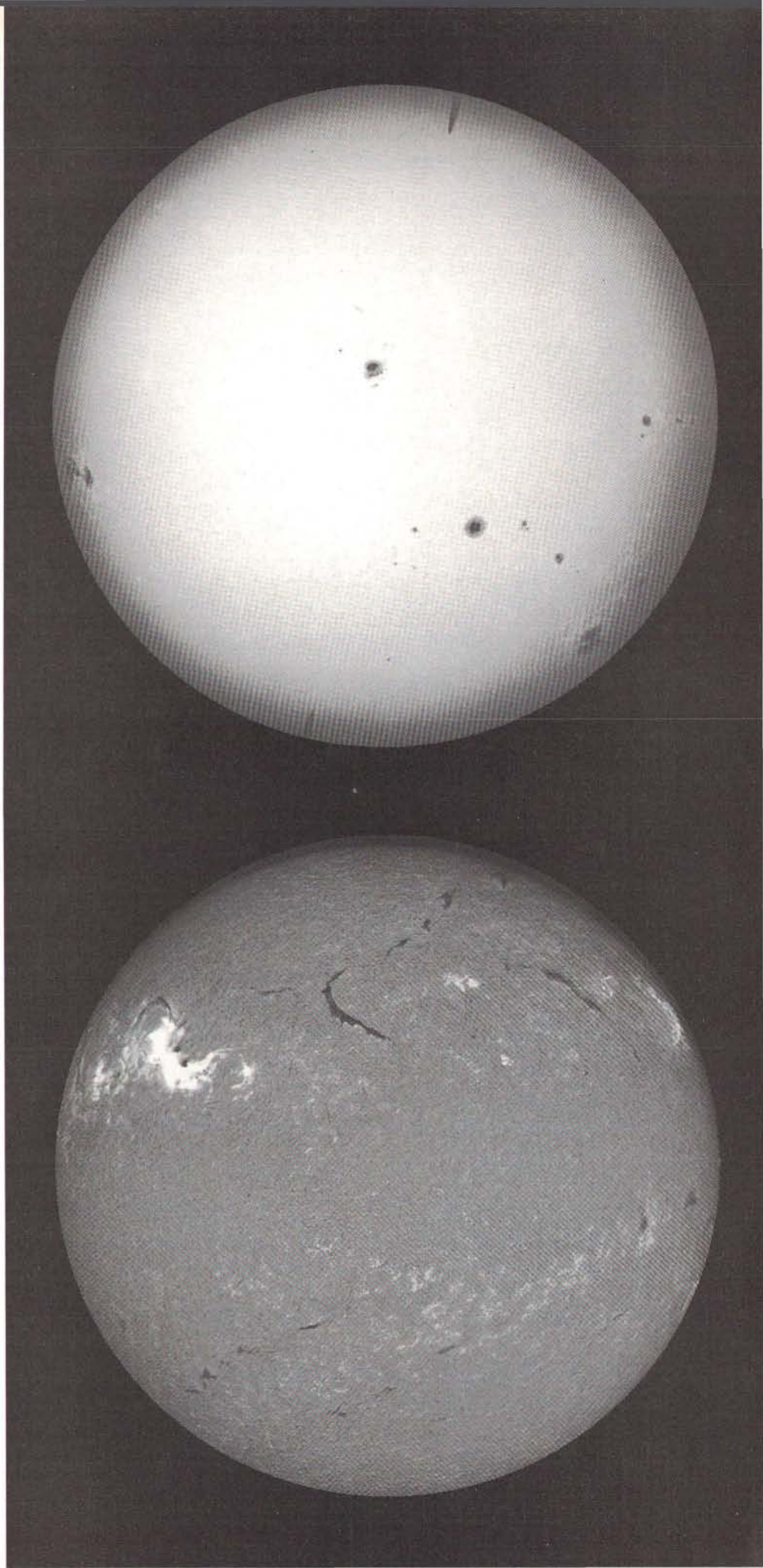
Recently, though, John A. Eddy, a solar astronomer at the University Corporation for Atmospheric Research in Boulder, Colorado, has accumulated convincing evidence that the dearth of sunspots documented by Maunder was a real and rather remarkable aspect of solar behavior [see "The Case of the Missing Sunspots," by John A. Eddy; SCIENTIFIC AMERICAN, May, 1977]. The Maunder minimum occurred during the most severe portion of a period of unusually cold weather, known as the Little Ice Age, which extended roughly from the 16th to the 18th centuries; the possible connection between these events is intriguing but still speculative. The Maunder minimum and the small peaks in the solar cycles at the beginning of the 19th century may explain why more than two centuries passed between the first European sunspot observations—which were made around 1610, shortly after the invention of the telescope, by Galileo and others—and Schwabe's discovery of the 11-year cycle.

The 11-year variation of sunspot number is now known to be merely the most visible aspect of a profound oscillation of the sun's magnetic field that affects many other aspects of the sun's surface and atmosphere and possibly its deepest interior as well. George Ellery Hale and his collaborators at the Mount Wilson Observatory in California found the

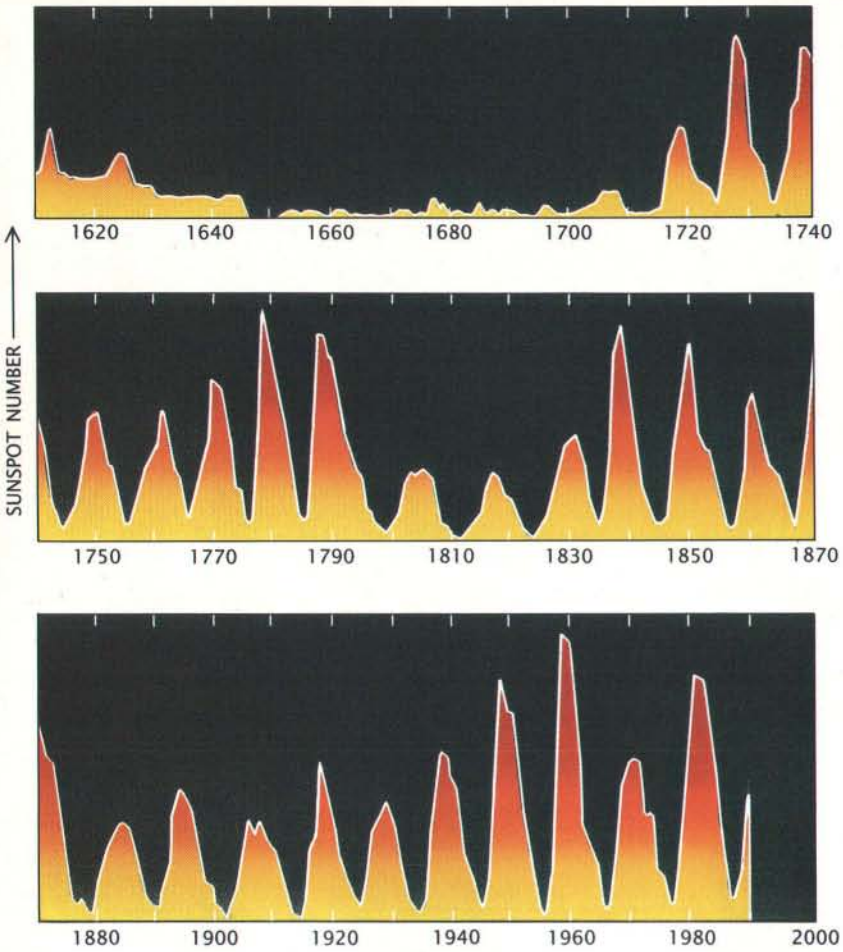
first evidence of solar magnetic oscillation in their measurements of sunspot spectra. They discovered that certain absorption lines in the spectra were broadened and polarized much like lines in laboratory spectra of magnetized gases, which had been studied by the Dutch physicist Pieter Zeeman. By analyzing this "Zeeman effect," they determined that the strength of the magnetic fields around sunspots is between 2,000 and 3,000 gauss, thousands of times stronger than the earth's field. They also showed that most spots occur in paired groupings that resemble giant magnetic dipoles (a bar magnet is one example of a dipole) and that are usually oriented roughly parallel to the solar equator.

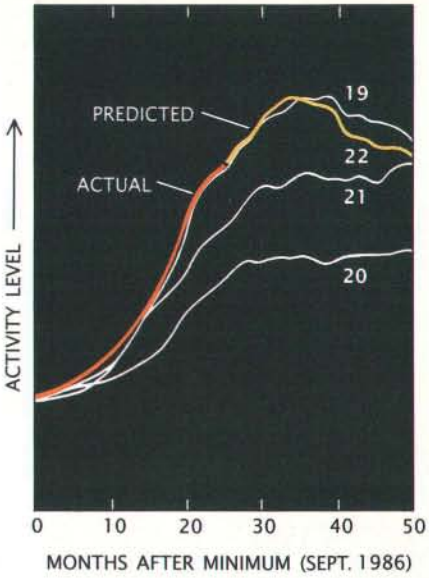
In 1912 Hale announced that the magnetic polarity of these sunspot dipoles had switched sign in the first spots of the new cycle that began that year. By 1924 he had collected enough observations to announce that this switch in polarity occurred at each activity minimum and was a basic feature of the sunspot cycle. He concluded that the 11-year cycle of sunspot number is actually half of a 22-year solar magnetic cycle, during which the polarity of sunspot groups reverses twice, hence returning to its original state.

Much more sensitive measurements of the sun's magnetic field are now made on a daily basis with the magnetograph, which was developed by Harold D. and Horace W. Babcock at Mount Wilson in 1951. One surprising finding extracted from such magnetograms, and from other evidence, is that the sun's surface magnetism is confined to small regions of intense magnetic field that cover only a few percent of the total area of the photosphere, the layer that forms the sun's visible surface. The largest areas of a single magnetic polarity are the sites of spot formation; smaller areas, ranging in size down to the resolution



TWO VIEWS OF THE SUN reveal complex processes beneath its mottled appearance. A visible-light photograph (*top*) shows sunspots and bright areas called faculae on the sun's "surface," or photosphere. Photographs made in a red wavelength emitted by hydrogen capture details in the hotter, higher-altitude region called the chromosphere (*bottom*). A bright flare, or solar eruption, can be seen. Dark filaments of relatively cool, dense gas are suspended by magnetic forces. Powerful magnetic fields determine much of the sun's outer structure.





SOLAR CYCLE manifests itself in the changing number of spots on the sun's visible surface (*left*). The dearth of spots between about 1645 and 1715, known as the Maunder minimum, appears to coincide with an era of unusually cold weather. The current cycle, cycle 22, is predicted to reach its peak this month; it has already surpassed the level of activity of the two previous cycles and may exceed cycle 19, the largest ever recorded (*above*).

limit of the most detailed magnetograms (about 200 kilometers), appear bright in most wavelengths of radiation. These especially luminous areas of the solar surface are called faculae (Latin for "little torches"); they were first seen early in the 17th century.

Hale and his collaborators tried inconclusively to observe the magnetic field near the sun's poles in an effort to determine if the sun had a global dipole field analogous to the earth's magnetic field. More sensitive measurements during the past 20 years indicate that the fields around the north and south solar poles are indeed usually, but not always, opposite in polarity and that they switch sign around the time of peak activity.

These observations also have revealed that the geometry of the sun's magnetic field is far more complicated than that of the earth's field, which can be reasonably modeled as a dipole magnet. The sun's field at low latitudes can be visualized as a series of field lines or magnetic tubes wrapped around the sun roughly parallel to its equator, submerged below the solar surface. Where these toruslike field lines emerge above the surface, they form looping stitches of magnetic field that extend into the outer layers of the solar atmosphere, sometimes reaching millions of kilometers out toward the planets before connecting back to the sun. Active regions, visible as spots and faculae, appear where these lines intersect the photosphere.

The mechanism that causes the solar magnetic cycle remains poorly understood, although it has been the focus of intense research during the past half-century. Astronomers generally agree that the observed changes in the sun's magnetism are caused by the motions of solar plasma forced across existing magnetic fields. Plasma is highly ionized gas—that is, gas in which many electrons have been stripped from their nuclei—and so is electrically conductive. Motions in the solar plasma induce both a current in the plasma and an associated magnetic field, which in turn intensifies the original field.

Unlike a solid body such as the earth, the sun's outer regions do not rotate at the same angular rate at all latitudes, a feature first demonstrated by the British astronomer Richard C. Carrington around 1860. The sun's equatorial regions complete one rotation in about 25 days, roughly 25 percent faster than the poles; there is a reasonably smooth variation in between. This differential rotation is probably a key factor in driving the dynamo that maintains the sun's magnetic field. A field line initially extending directly along the surface between the solar poles and constrained to move with the surface plasma would be progressively stretched by the faster equatorial rotation. After a few solar rotations, the line would be wound almost parallel to the equator. This deformation of the field lines probably accounts for the geometry of the solar magnetic field and for the east-west orientation of sunspot groups [*see illustration on page 30*]. The stretching of magnetic field lines increases their intensity to the high values measured in sunspots.

The eruption of magnetic flux from the sun is believed to be partially responsible for the change in polarity of the sunspot fields between cycles. As the magnetic tubes that give rise to active regions emerge from the sun's interior, their magnetic flux eventually disperses across the solar surface. At

the same time, the flux is pushed outward into higher layers of the solar atmosphere by new magnetic fields, which rise from below as a result of the sun's differential rotation. The result is a shedding of the old flux that somehow removes the original polarity and leaves the new flux with a net excess of the opposite polarity.

The combination of the shedding of "old" magnetic field by eruption and dispersal and the generation of new field by differential rotation is probably sufficient to explain the solar magnetic cycle, but little is known about the process by which the sun rids itself of the old polarity. Computer simulations of the dynamics of solar-plasma motions and of their interactions with the magnetic field have difficulty producing the 11-year cycle in conjunction with the sun's differential rotation. A major uncertainty has been the nature of the sun's internal rotation, which was essentially unknown until recently. Observations of the sun's global oscillations [see "Helioseismology," by John W. Leibacher, Robert W. Noyes, Juri Toomre and Roger K. Ulrich; SCIENTIFIC AMERICAN, September, 1985] are providing a window into the interior of the sun, making it possible to analyze the depth profile of solar rotation and to calculate its influence on the solar dynamo. Recent findings indicate that the outer 30 percent of the sun's interior rotates differentially, much as its surface does. This suggests that much of the dynamo action in the sun may take place far below the surface.

Although astronomers still remain far from a comprehensive understanding of the sun's magnetic oscillations, measurement of the magnetic cycle's effects on key solar outputs would in itself be an important advance. Climatologists would be quite satisfied to know the amplitude and characteristic time scales of solar-luminosity changes even if the astrophysical explanation of the changes had to wait. Fortunately, substantial progress has been made in the past few years in this kind of empirical understanding of the sun's behavior.

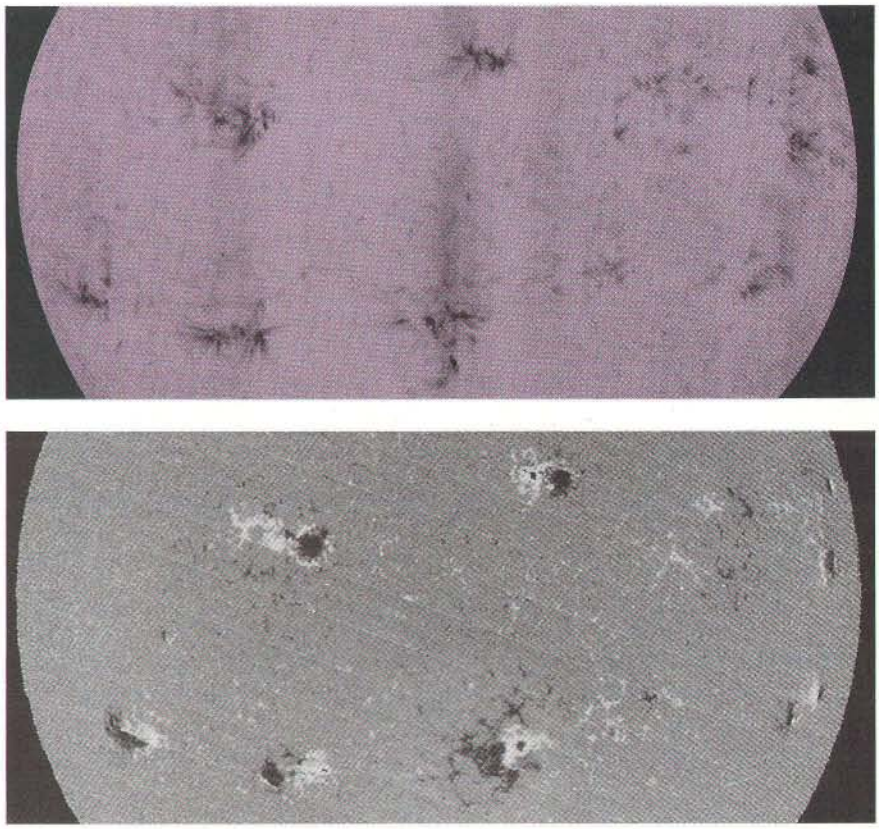
One important advance has been the discovery of cyclic variations in the sun's total light output, known as the total solar irradiance, or—ironically—the solar "constant." The vagaries of the earth's atmosphere have made measurement of the solar constant notoriously difficult, but satellite instruments now reveal that it varies by as much as .2 percent over time scales of weeks. This relatively short-term variation is caused by the passage of dark sunspots and bright faculae across the solar disk as the sun executes its approximately monthly rotation.

These short-term fluctuations were first identified clearly—in data obtained in 1980 with radiometers flown on the *Solar Maximum Mission* (*SMM*) and *Nimbus 7* satellites—by Richard C. Willson of the Jet Propulsion Laboratory in Pasadena, Calif., and John H. Hickey of the Eppley Laboratory in Newport, R. I., respectively. It was harder to identify longer-term changes in the solar constant over the course of the sunspot cycle, because these variations have turned out to be on the order of .1 percent; slow changes in radiometer calibration at this level were difficult to rule out. Convincing evidence finally arrived when readings from both radiometers, which had been falling along with the sun's decreasing activity level since 1980, flattened out in 1986 and began to rise in 1987 as the sun moved into its current cycle and increased its activity.

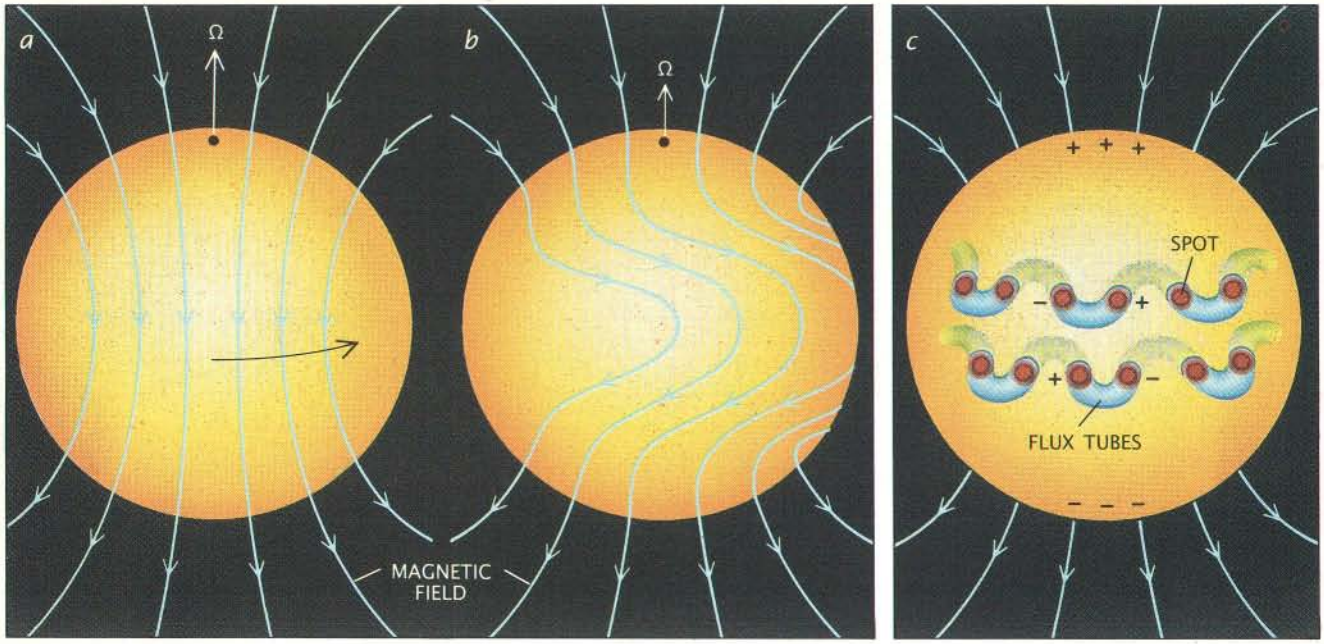
Data from the two satellites indicate that the sun's brightness decreased by about .1 percent between the peak of solar activity in 1981 and its minimum in mid-1986. Surprisingly, the sun grew more luminous as the number of sunspots on its surface grew larger. Analysis of this behavior by Judith L. Lean of the Naval Research Laboratory and myself indicates that the increase in photospheric area covered by bright faculae outweighs the increase in area of dark spots as solar activity increases.

Do these changes in the solar constant affect the earth's weather or climate? It is fairly simple to calculate the effect of variations of solar irradiance on the earth's global mean temperature. Present-day climate models suggest that the size of the effect is well below .1 kelvin. This is a small fraction of the global-warming effect (typically several tenths of a degree) that is expected from the measured increase in the concentration of atmospheric carbon dioxide during the past few decades.

The measurements of solar irradiance made from space so far extend



MAGNETIC FIELD LINES trap huge streamers of hot, ionized gas in the solar atmosphere, or corona. These appear as dark lines in photographs of the sun made in short-wavelength ultraviolet light (*top*). The largest streamers emanate from sunspots. A magnetic image of the sun made at the same time reveals the intense magnetic fields associated with the streamers (*bottom*). Dark and light areas represent opposite magnetic polarities; the strongest fields occur in bipolar active regions.



DIFFERENTIAL ROTATION of the sun is believed to drive the solar magnetic cycle. The solar surface drags existing magnetic field lines along with it as the sun rotates. A complete rotation requires about 25 days near the sun's equator and about 28 days at midlatitudes; over several rotations, a field line that initially followed a straight north-south path (*a*) winds up, stretches horizontally and intensifies (*b*). Sunspots appear where the most intense tubes of magnetic flux emerge at the surface; the points of emergence and reentry of the tubes have opposite polarities (*c*). Sunspots usually form in pairs at similar latitudes because they follow the horizontal stretching of the magnetic field near the surface. Sunspot pairs are huge magnetic dipoles; their orientation in the Northern Hemisphere is opposite to that in the Southern Hemisphere.

over barely one sunspot cycle. Current information is insufficient to tell whether larger solar-irradiance variations occur, perhaps in conjunction with slower variations in solar activity such as the one that produced the Maunder minimum. It has been estimated that during the Little Ice Age global mean temperatures dipped by as much as about .5 kelvin below the long-term average, enough to precipitate a period of significant glacial advance and to cause a series of failed crops in Europe. If the cooling were a result of a change in the sun's luminosity, it would have required a dip in the solar irradiance of between .2 and .5 percent acting over several decades, according to calculations based on standard climate models. A dedicated program of highly precise irradiance monitoring from outer space could detect long-term changes in the sun's luminosity and enable astronomers to determine whether a future lull in solar activity is likely to lead to another extended period of global cooling.

Over the years numerous attempts have been made to find links between the solar cycle and terrestrial weather. The eminent British astronomer Sir William Herschel guessed (correctly!) that the sun is brightest during sunspot maxima, and he mused that the resulting higher temperatures would improve the wheat crop and cause prices to drop. In 1801 he announced that the price of wheat was indeed correlated with the sunspot cycle. The correlation vanished, however, and Herschel turned his attention to other matters. Many other apparent "links" have been similarly short-lived, and all suffer from the fact that they are statistical rather than causal connections; nobody has yet demonstrated a plausible mechanism whereby such tiny variations in the solar constant could have an appreciable effect at the earth's surface.

Yet the quest persists. In 1987 Karin Labitzke of the Free University of Berlin reported the most convincing link yet found. She discovered that the occurrence of mid-winter warmings in the U.S. and Western Europe has correlated remarkably well with the solar cycle over the past 40 years, provided the switch in direction of stratospheric winds roughly every two years is taken into account. The relation she found has withstood repeated statistical tests, and it correctly predicted the warming that accounted for the very mild winter of 1988-89 in the U.K. and Western Europe. An unambiguous and physically explicable connection between solar and climatic variability would represent a tremendous advance in the understanding of the relation between the earth and its star.

The scope of solar variability extends far beyond changes in the appearance and brightness of the photosphere. The magnetic cycle also affects the progressively higher layers of the solar atmosphere, known as the chromosphere, corona and solar wind. The plasma temperature in these regions actually exceeds that of the photosphere even though they are farther from the sun's nuclear heat source. Because the energy losses from these tenuous plasmas are very low, extremely high temperatures—up to millions of degrees—can be maintained by relatively small inputs of energy. This energy is probably derived from the dissipation of sound waves produced by the churning of the photosphere and of electric currents associated with magnetic fields at or below the photosphere.

These hot outer layers of the sun's atmosphere are responsible for the sun's highly variable emissions of X rays and of extreme ultraviolet radiation (EUV), or the wavelengths between about 100 and 1,000 angstrom units (one angstrom unit is 10^{-10} meters). The chromosphere also emits a substantial fraction of the sun's ultraviolet radiation at wavelengths between about 1,600 and 3,200 angstroms and

probably accounts for most of the variability in these radiations. Solar X rays have less impact on terrestrial life than UV and EUV emissions, which are of concern because they significantly affect the earth's atmosphere.

The cause of the variability in EUV radiation seems clear from analysis of observations obtained from solar telescopes on *Skylab* in 1973–74 and from more recent satellites such as *SMM*. Intense, locally closed dipole magnetic fields in active regions act as magnetic "cages" that prevent the escape of hot coronal plasma from the sun's gravitational attraction. This trapped plasma is about 10 times denser than that in the quiet surrounding areas, and the denser plasma radiates much more intensely in the EUV. Active regions are therefore the major source of solar EUV emissions, and these emissions rise and fall along with the changes in the solar-activity cycle.

The sun's EUV output variation over the course of an entire 11-year sunspot cycle has not yet been measured successfully. It is a challenge to design spectrometers and detectors for this spectral range and to ensure that their calibration will remain sufficiently steady under powerful EUV irradiation in space to identify reliably slow changes of even several tens of percent. Measurements in the prominent hydrogen emission line at 1,216 angstroms (the Lyman alpha line) indicate a variation in intensity of about a factor of two.

The rapid rise of the current solar cycle has produced a dramatic increase in the solar EUV flux, raising concerns at the National Aeronautics and Space Administration about the orbital lifetimes of the Hubble Space Telescope and the Long Duration Exposure Facility (LDEF), a massive orbiting platform containing 57 zero-gravity experiments. The problem is atmospheric drag: during periods of high solar activity, the increased EUV heating of the earth's upper atmosphere above about 100 kilometers can cause temperatures in the ionosphere to soar to almost three times the values encountered at periods of low activity. These higher temperatures in turn enable the atmosphere to support a gas density as much as 50 times higher at an altitude of 600 kilometers, where the space telescope is supposed to orbit. The greater atmospheric density would lead to increased drag and a much shorter orbital lifetime before a reboost by the space shuttle would be necessary for the $1-billion telescope. NASA has therefore asked solar astronomers to predict both the magnitude of the current cycle's activity and its time of maximum in an effort to choose the best launch date for the space telescope and to plan for the recent shuttle retrieval of the LDEF.

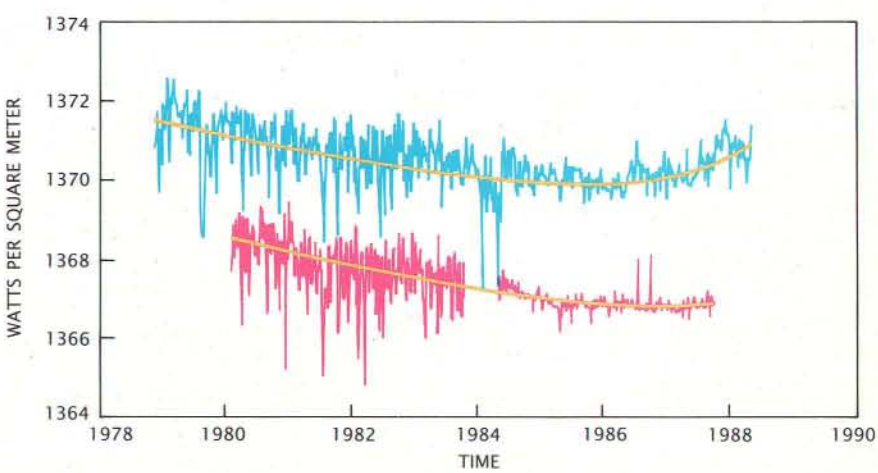
Measuring the sun's 11-year variation is somewhat easier in the ultraviolet than it is in the EUV, and during the past decade more effort has been devoted to observations in ultraviolet wavelengths because they directly affect the earth's ozone layer. Data from the *Nimbus 7* and *Solar Mesosphere Explorer* satellites clearly show a 27-day variation (caused by the sun's rotation) at wavelengths shorter than about 3,000 angstroms. The amplitude of the 11-year solar-cycle variation in the UV is elusive because of calibration difficulties, but it seems to range from as much as 20 percent around 1,500 angstroms to only 1 or 2 percent at wavelengths longer than 2,500 angstroms.

Current models of ozone production indicate that these changes in solar UV emissions might cause a variation of from 1 to 2 percent in total global ozone. This could account for much of the global decrease in stratospheric ozone measured by satellites between 1978 and 1985, a period of mostly declining solar activity. The effects of the solar cycle clearly must be taken into account in attempts to identify the cause of this ozone decrease and the longer-term decrease measured by ground-based instruments between 1969 and 1986. The slow decrease in global ozone is less dramatic than the more recently identified hole in the ozone layer over the Antarctic, but, if this decrease continues, it will be even more serious.

The sun's output of charged particles, which depends primarily on conditions in the layers above the photosphere, also changes over the course of the solar cycle. The most important of these particles in terms of their impact on terrestrial systems are the high-energy protons that are occasionally spewed out by explosions in the solar corona. The earth is also affected by the more general outpouring of coronal plasma called the solar wind.

The high-energy solar protons observed at the earth range in energy from about 10 million to 10 billion electron volts (for comparison, a photon of visible light has an energy of about two electron volts). The most energetic protons travel at close to the speed of light and arrive at the earth about eight minutes after certain of the largest solar flares have occurred. These flares are huge eruptions in solar-active regions; they cause these regions to grow dramatically brighter in X rays and in EUV. Flares are thought to derive their energy from the rapid annihilation of the intense magnetic fields, which heats the plasma and produces powerful electric fields that accelerate charged particles.

Large proton events are of concern to commercial airlines, particularly those flying along polar routes, where the earth's magnetic field lines curve



FLICKERING of the sun was recorded by radiometers on two satellites, *Nimbus 7* (*blue*) and *Solar Maximum Mission* (*red*). Short-term decreases in solar output produced the sharp spikes in the *SMM* data, and most of those seen in the *Nimbus 7* data, which also included some instrument noise. On the average (*yellow line*), the sun shone brightest at the time of maximum sunspot activity. Apparently the greater number of bright faculae at maximum activity outweighed the effect of dark spots.

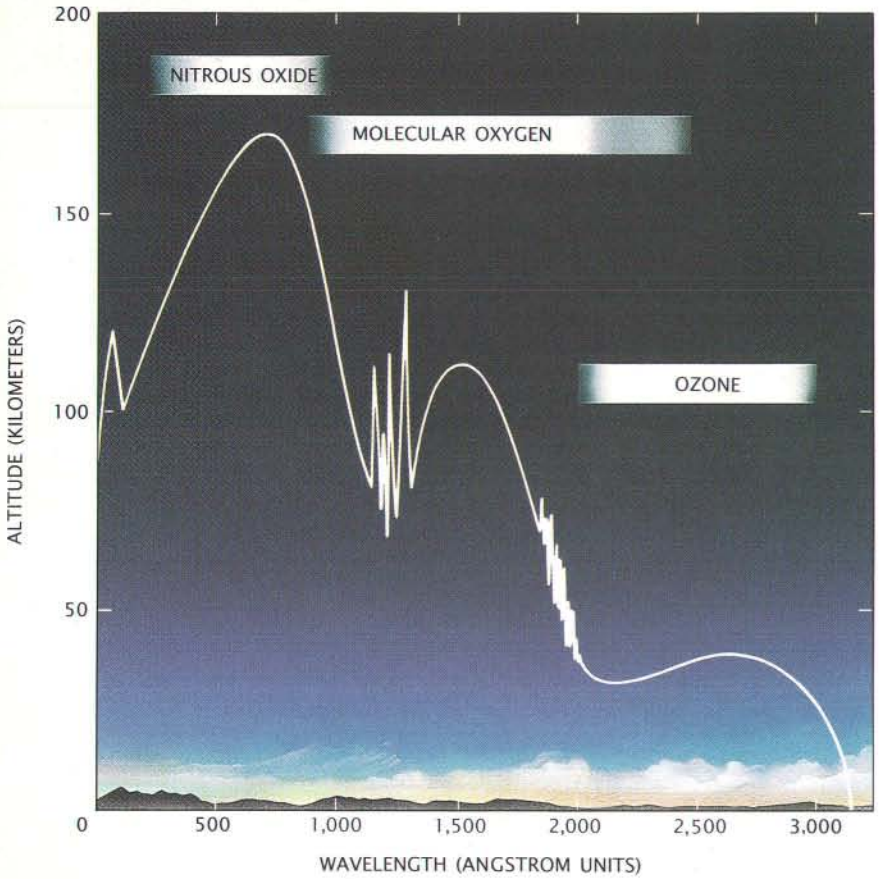
down to the surface and allow charged particles to penetrate to low altitudes, exposing passengers to elevated levels of radiation. These events pose a more serious threat to astronauts, especially those who may fly on polar-orbiting satellites. Proton events also have been implicated in the failure of computer systems; last August one such event temporarily shut down the Toronto stock exchange. Only a few dozen such large flares occur during a solar cycle, but their frequency is much higher near sunspot maxima than it is near minima.

Variations in the continuous flow of solar-wind plasma past the earth give rise to a quite different class of interactions. This relatively low-energy plasma can be visualized as the "overflow" of the solar corona, which is too hot to be entirely contained by the inward pull of the sun's gravitational field. The solar wind is excluded from the immediate vicinity of the earth by the planet's magnetic field, which exerts an electromagnetic force on charged particles attempting to cross the earth's field lines; the volume around the earth from which the bulk of the solar wind is excluded is called the magnetosphere. Flares and other magnetic eruptions in the solar atmosphere lead to disruptions in the solar wind that alter the plasma pressure exerted on the magnetosphere.

The resulting fluctuations in the geomagnetic field are typically only .1 percent of its roughly one-gauss intensity. But the electric currents they can induce in large-scale conductors at the earth's surface, such as power-line networks and oil pipelines, can have dramatic effects. For instance, on March 13, 1989, a powerful geomagnetic storm caused by flares associated with one of the largest spots ever observed knocked out electricity throughout the province of Quebec.

Such potent geomagnetic storms are caused in part by the flares that erupt in active regions of the sun, and so these storms increase in frequency along with sunspot number during the magnetic cycle. The steadier outflow of the solar wind seems to originate from areas of the corona outside active regions, where the sun's magnetic field lines stretch outward to the earth and beyond, thereby creating a path along which charged particles can travel relatively unencumbered.

In some regions of the sun, the open field-line configuration allows charged particles in the solar wind to escape quite easily. This results in areas of depleted coronal plasma, or "coronal holes." Holes are always present near the solar poles, but they can also evolve at lower latitudes. Low-latitude holes produce high-velocity solar-wind streams that can spray the earth directly and recurrently as the sun rotates. The appearance of holes is linked to the solar cycle, but with a different sort of modulation than is observed for sunspots. Although the data are limited to less than two full solar cycles, it appears that the largest low-latitude holes develop during the declining phase of a cycle, so that their contribution to geomagnetic activity is greatest a few years after the peak sunspot number.



PENETRATION DEPTH of solar radiation through the earth's atmosphere varies according to its wavelength. The graph shows the altitudes at which about half of a given radiation is absorbed. Fortunately for life, nitrous oxide in the thin atmosphere more than 50 kilometers above the earth's surface blocks the sun's highly variable, short-wavelength ultraviolet emissions. At lower altitudes, ozone and molecular oxygen absorb longer-wavelength ultraviolet rays, which are also harmful to life. Changes in the sun's ultraviolet emissions affect the structure of the ozone layer.

The many ways in which solar variability affects the terrestrial environment underscore how useful it would be to be able to predict the size and timing of the next sunspot maximum. Present attempts at prediction are extremely limited because they rely on empirical rules derived from the behavior of past cycles. Nevertheless, these rules have provided some useful estimates of the current cycle—cycle 22—and have been useful for NASA's calculations of orbital lifetimes. It appears from current behavior that this cycle will equal or exceed the activity level of the largest cycle reliably recorded: cycle 19, which peaked in 1957.

The ability to predict future solar activity depends critically on knowledge of past activity. Several sources of information stretch farther into the past than the first telescopic observations of spots in 1610. Naked-eye observations of large sunspots extend back to at least the fourth century B.C., although before Galileo's time, spots were usually thought to be planets or other nonsolar phenomena crossing the solar disk. Auroras visible at low latitudes are caused primarily by solar flares; aurora observations have proved to be a less ambiguous means of inferring past solar behavior.

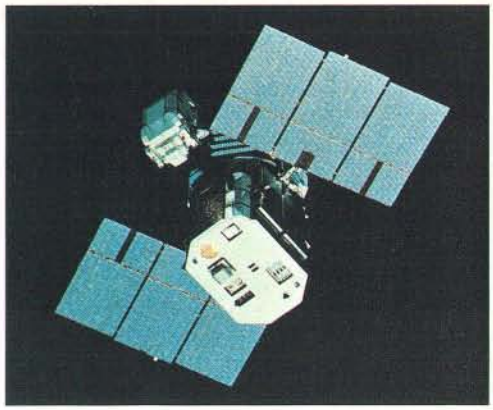
A particularly long record of solar activity is hidden in the historical abundances of carbon 14, a radioac-

tive isotope of ordinary carbon 12. The production of C-14 in the earth's atmosphere is determined by the flux of high-energy particles known as galactic cosmic rays, which are produced by energetic processes outside the solar system. The ability of these cosmic rays to penetrate into the solar system is decreased by the strength and geometry of magnetic fields carried out from the sun by the solar wind at high solar-activity levels. In the course of photosynthesis, plants take up C-14 along with other isotopes of carbon and incorporate it into their structure. Solar-activity levels over the past two millennia can be estimated by studying the relative abundance of C-14 in the tree rings of old, but still living, trees; the ages of the rings can be calculated simply by counting back from the present.

The results of ancient spot observations, auroral sightings and C-14 data were brought together in 1976 in Eddy's groundbreaking study. Eddy determined beyond a reasonable doubt that the Maunder minimum coincided with dramatic reductions in the level of solar activity, as indicated by the paucity of auroras and by high C-14 levels. He and others have shown further that such episodes of abnormally low solar activity lasting many decades are a fairly common aspect of solar behavior—another episode, the Spörer minimum, occurred between about 1450 and 1550. On the other hand, an extended period of high solar activity between about 1100 and 1250 coincided with relatively warm weather that seems to have made Viking migrations to Greenland and the New World possible. The historical record indicates that another lull in solar activity might reasonably be expected in the next century.

Another source of evidence regarding the "normal" behavior of the solar cycle, and of the sun's variability in the distant past, comes from observations of other stars similar in mass to the sun. In a pioneering study published in 1976, Olin C. Wilson of the Mount Wilson Observatory showed that chromospheric emissions observable in visible light from many such stars show cyclic variations over periods similar to that of the solar cycle. This provided the first strong evidence that magnetic-activity cycles are a common feature of stars resembling the sun in mass and age (and so also in size and temperature). Measurements of stellar-light curves are also providing tentative indications of differential rotation in these stars.

Studies of stars much younger than the sun show how intense and variable the sun's ultraviolet and total light outputs might have been billions of years ago, when life first appeared on the earth. These studies reveal that the level of a star's photospheric and chromospheric activity, and of its coronal X-ray emission, correlates quite closely with its rate of rotation. Younger stars generally rotate faster than older ones and therefore are more intense and variable in UV light and X rays; they also tend to vary more in total luminosity. One recent and unexpected finding is that younger stars seem to be fainter during periods of high activity, implying that, for them, the effect of "starspots" overwhelms that of faculae—the reverse of the case for the sun.



SOLAR TELESCOPES have grown increasingly complex. In the 1850's J. Rudolf Wolf made the first systematic measurements of the sunspot cycle using a small refracting telescope that is still operating at the Zurich Observatory (*left*). In the 1980's the *Solar Maximum Mission* satellite (*right*) performed delicate measurements of the sun's atmospheric structure and variable radiations. The satellite crashed to the earth on December 2, 1989; its orbit decayed because of the expansion of the earth's upper atmosphere caused by the high levels of solar activity that it was monitoring.

Is it realistic to expect that astronomers will ever be able to make long-term predictions about the behavior of the sun? There may be fundamental restrictions on the possibility of such predictions if the processes driving the solar cycle are nonlinear. Nonlinear systems do not behave in the predictable manner of simple oscillators, such as a pendulum; relatively straightforward feedback of an "effect" on its "cause" can lead to bewilderingly complex behavior. Even when their behavior is governed by a well-understood set of forces, nonlinear oscillators can be so sensitive to initial conditions that predictions extending substantially into the future become impossible.

In the 1950's, while he was studying long-range weather forecasting, Edward N. Lorenz of the Massachusetts Institute of Technology made important contributions to the study of nonlinear behavior. He showed that feedback between various mechanisms in the atmosphere makes prediction of future weather conditions difficult because current pressure, temperature and winds must be known with impossible accuracy in order to predict precisely more than several days into the future. Analysis of the dynamics of the solar cycle suggests that its fairly regular behavior between about 1700 and the present as well as its disappearance for about six 11-year periods between 1645 and 1715 may be characteristic of a nonlinear oscillator.

Research now under way should help establish whether the solar-activity cycle is predictable, at least in principle, or whether it is chaotic. Even if the solar cycle is unpredictable, comprehending the possible relations between slow changes in solar activity and climate will be important in unraveling the earth's past climatic record—and in preparing us for variations that can be expected in the centuries to come.

FURTHER READING

SUN, WEATHER, AND CLIMATE. J. R. Herman and R. A. Goldberg. National Technical Information Service, NASA-SP-426; 1978.

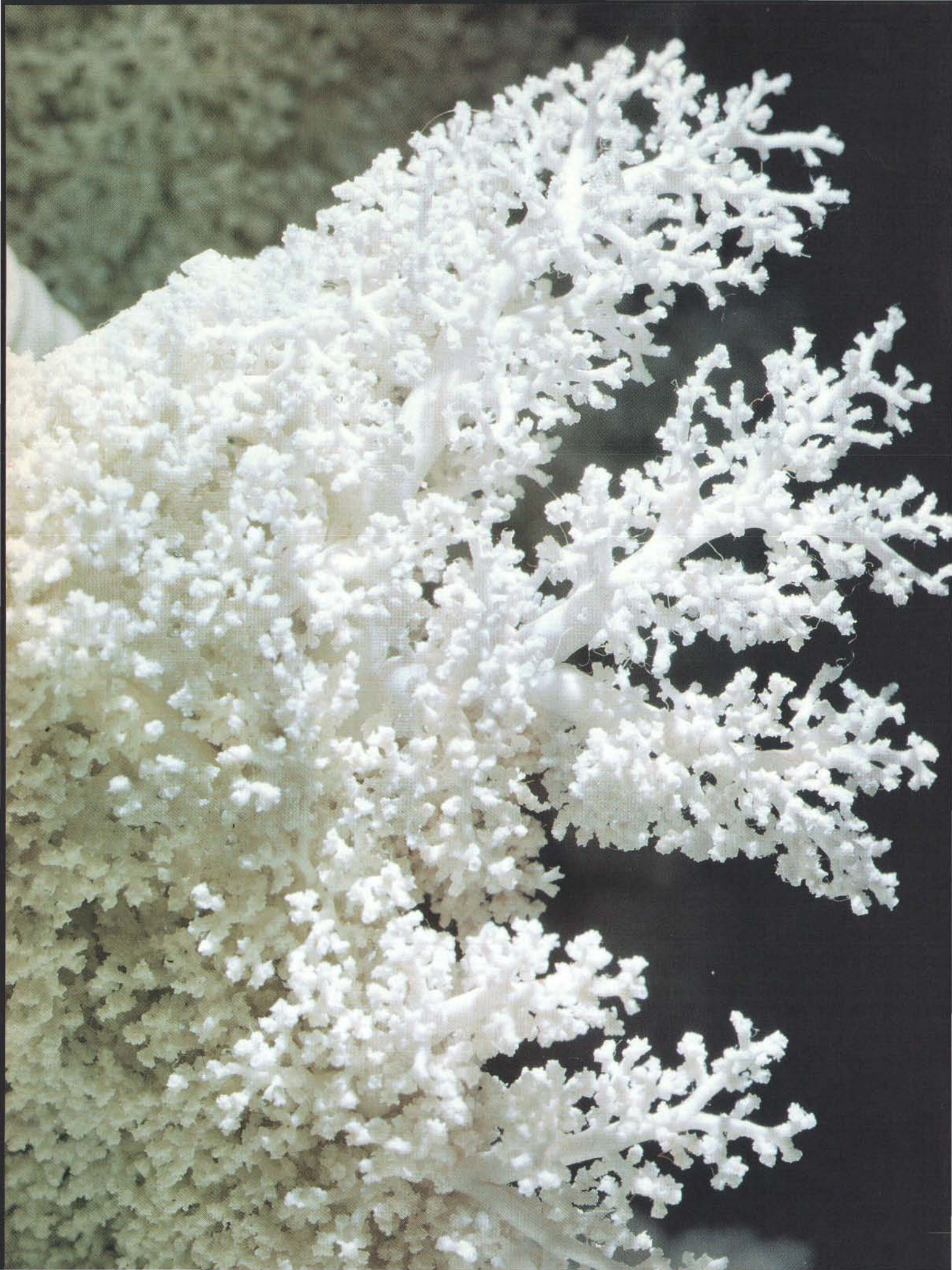
THE SUN, OUR STAR. Robert W. Noyes. Harvard University Press, 1982.

SUN AND EARTH. Herbert Friedman. Scientific American Books, Inc., 1986.

ASTROPHYSICS OF THE SUN. Harold Zirin. Cambridge University Press, 1988.

THE RESTLESS SUN. Donat G. Wentzel. Smithsonian Institution Press, 1989.

SOLAR ASTROPHYSICS. Peter Foukal. Wiley Interscience, 1990.



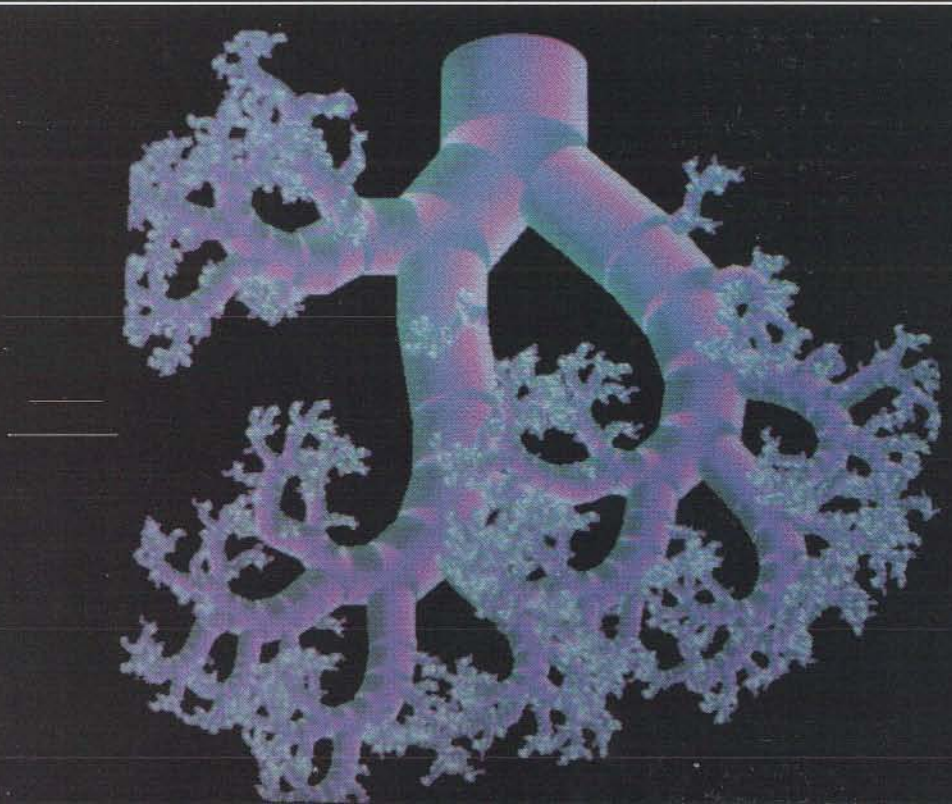
Chaos and Fractals in Human Physiology

Chaos in bodily functioning signals health. Periodic behavior can foreshadow disease

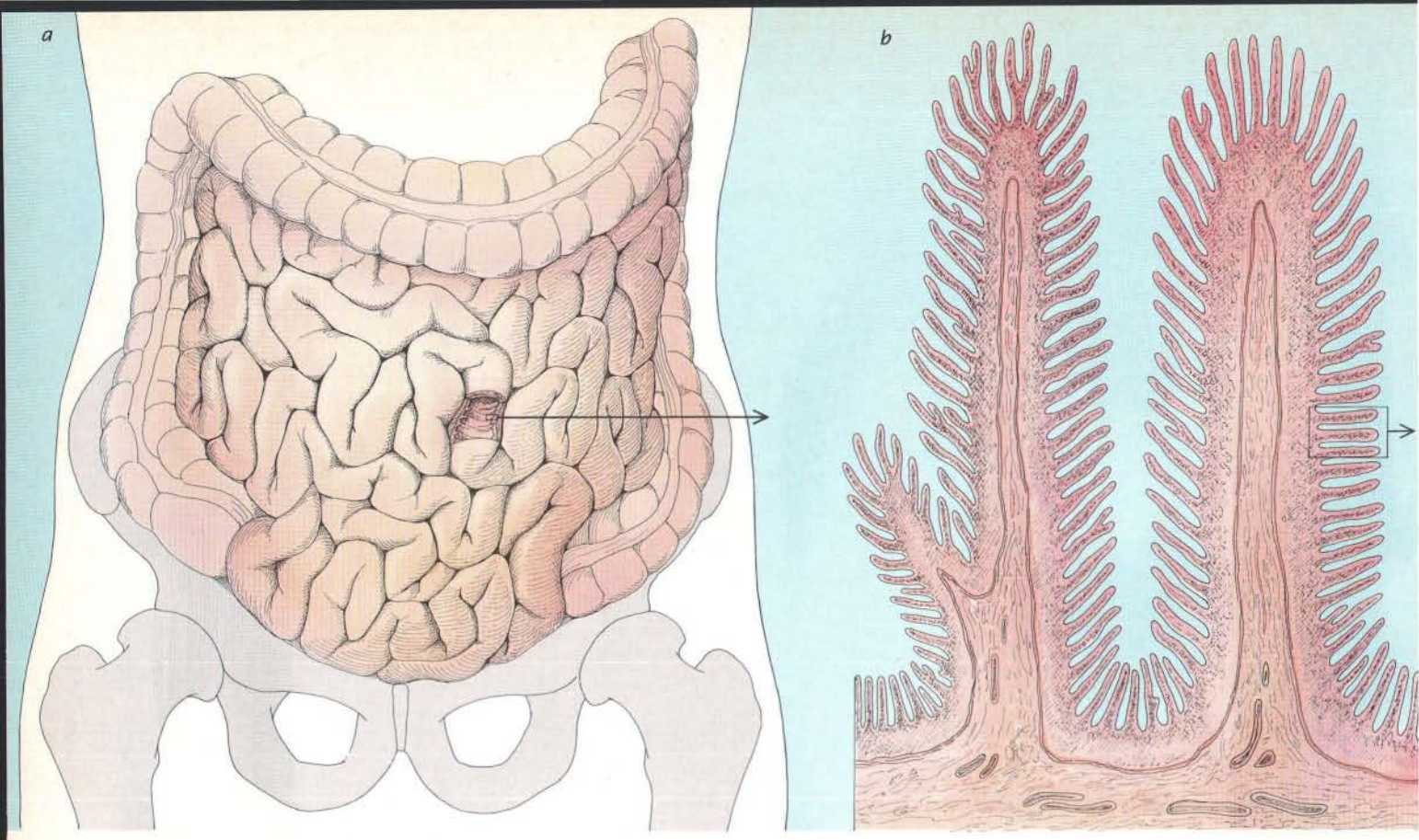
by Ary L. Goldberger, David R. Rigney and Bruce J. West

A medical student monitoring the rhythms of a heart notices that the tempo sometimes changes dramatically from minute to minute and hour to hour. A clinician maneuvering a bronchoscope into a lung observes that the trachea branches into smaller and smaller airways. The student senses that the interval between heartbeats varies chaotically. Perhaps the clinician recognizes that the network of airways resembles a fractal. Physiologists and physicians have only recently begun to quantify such possibilities of chaotic dynamics and fractal architectures. Their investigations are challenging long-held principles of medicine and are revealing possible forewarnings of disease.

The conventional wisdom in medicine holds that disease and aging arise from stress on an otherwise orderly and machinelike system—that the stress decreases order by provoking erratic responses or by upsetting the body's normal periodic rhythms. In



AIRWAYS OF THE LUNG **(left)** ***shaped by evolution and embryonic development resemble fractals generated by computer*** **(below).** ***The bronchi and bronchioles of the lung*** **(here a rubber cast)** ***form a "tree" that has multiple generations of branchings. The small-scale branching of the airways looks like branching at larger scales. When physiologists quantified observations of the branching pattern, they discovered that the lung tree has fractal geometry.***



the past five years or so we and our colleagues have discovered that the heart and other physiological systems may behave most erratically when they are young and healthy. Counterintuitively, increasingly regular behavior sometimes accompanies aging and disease.

Irregularity and unpredictability, then, are important features of health. On the other hand, decreased variability and accentuated periodicities are associated with disease. Motivated by these ideas, we and other physiologists have looked for periodic behavior that might indicate developing sickness (especially diseases of the heart). In addition, we have begun to analyze the flexibility and strength of irregular fractal structures and the adaptability and robustness of systems that exhibit apparently chaotic behavior.

Chaos and fractals are subjects associated with the discipline of nonlinear dynamics: the study of systems that respond disproportionately to stimuli. The theory of nonlinear dynamics provides insights into the phenomenon of epidemics, the kinetics of certain chemical reactions and the changes in the weather. Under some circumstances deterministic nonlinear systems—those that have only a few simple elements—behave erratically, a state called chaos. The deterministic chaos of nonlinear dynamics is not the same as chaos in the dictionary sense of complete disorganization or randomness. Nonlinear chaos refers to a constrained kind of randomness, which, remarkably, may be associated with fractal geometry.

Fractal structures are often the remnants of chaotic nonlinear dynamics. Wherever a chaotic process has shaped an environment (the seashore, the atmosphere, a geologic fault), fractals are likely to be left behind (coastlines, clouds, rock formations). Yet at first the mathematics of fractals developed independently of nonlinear dynamics, and even today the connections between the disciplines are not fully established.

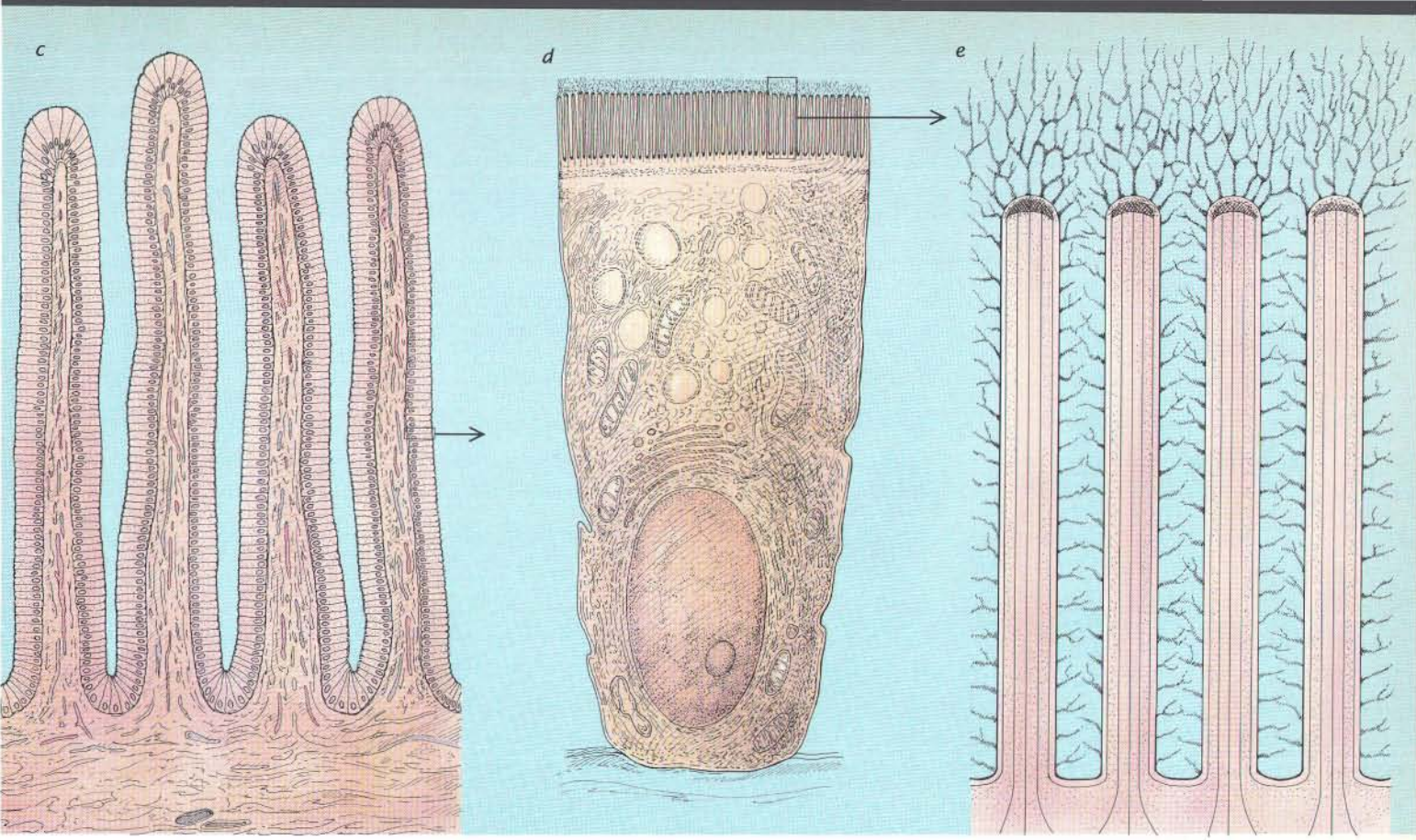
A fractal, as first conceived by Benoit B. Mandelbrot of the IBM T. J. Watson Research Center, consists of geometric fragments of varying size and orientation but similar shape. Certain neurons, for instance, have a fractallike structure. If one examines such neurons through a low-power microscope lens, one can discern asymmetric branches, called dendrites, connected to the cell bodies. At slightly higher magnification, one observes smaller branches on the larger ones. At even higher magnification, one sees another level of detail: branches on branches on branches. Although at some level the branching of a neuron stops, idealized fractals have infinite detail.

Perhaps it is even more remarkable that the details of a fractal at a certain scale are similar (though not necessarily identical) to those of the structure seen at larger or smaller scales. If one saw two photographs of the dendrites at two different magnifications (without any other reference), one would have difficulty in deciding which photograph corresponded to which magnification. All fractals have this internal, look-alike property called self-similarity.

Because a fractal is composed of similar structures of ever finer detail, its length is not well defined. If one attempts to measure the length of a fractal with a given ruler, some details will always be finer than the ruler can possibly measure. As the resolution of the measuring instrument increases, therefore, the length of a fractal grows.

Because length is not a meaningful concept for fractals, mathematicians calculate the "dimension" of a fractal to quantify how it fills space. The familiar concept of dimension applies to the objects of classical, or Euclidean, geometry.

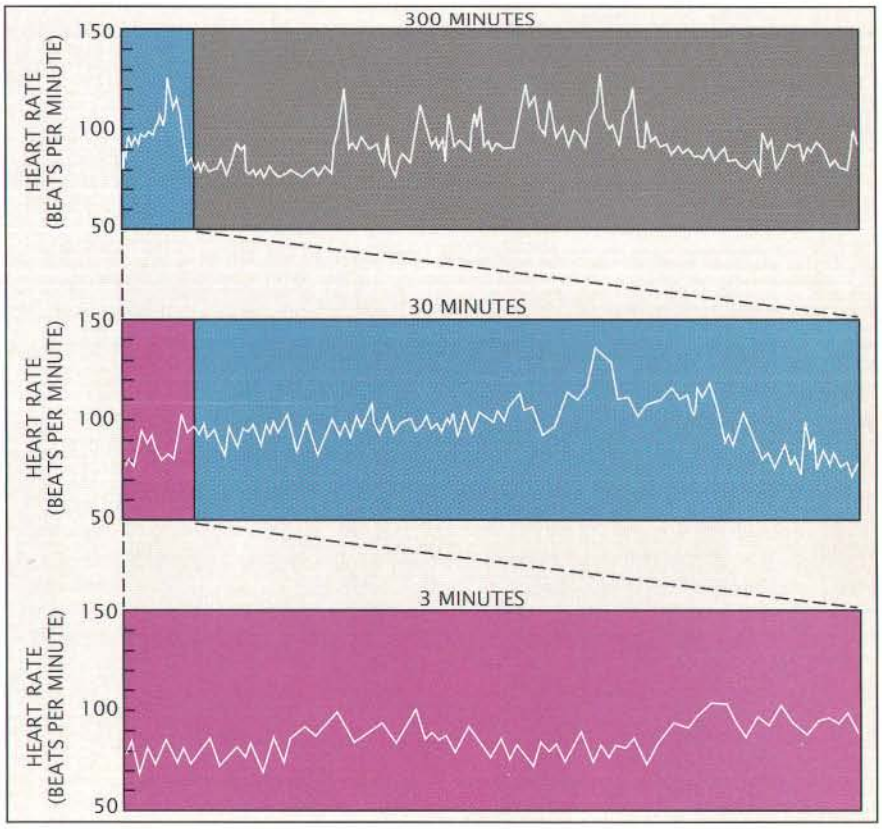
ARY L. GOLDBERGER, DAVID R. RIGNEY and BRUCE J. WEST have collaborated on studies of nonlinear dynamics in physiology. Goldberger is assistant professor of medicine at Harvard Medical School and co-director of the Electrocardiography and Arrhythmia Laboratories at Beth Israel Hospital, Boston. Rigney is assistant professor of medicine at Harvard and research affiliate at the Massachusetts Institute of Technology. West is professor of physics and chair of the department of physics at the University of North Texas.

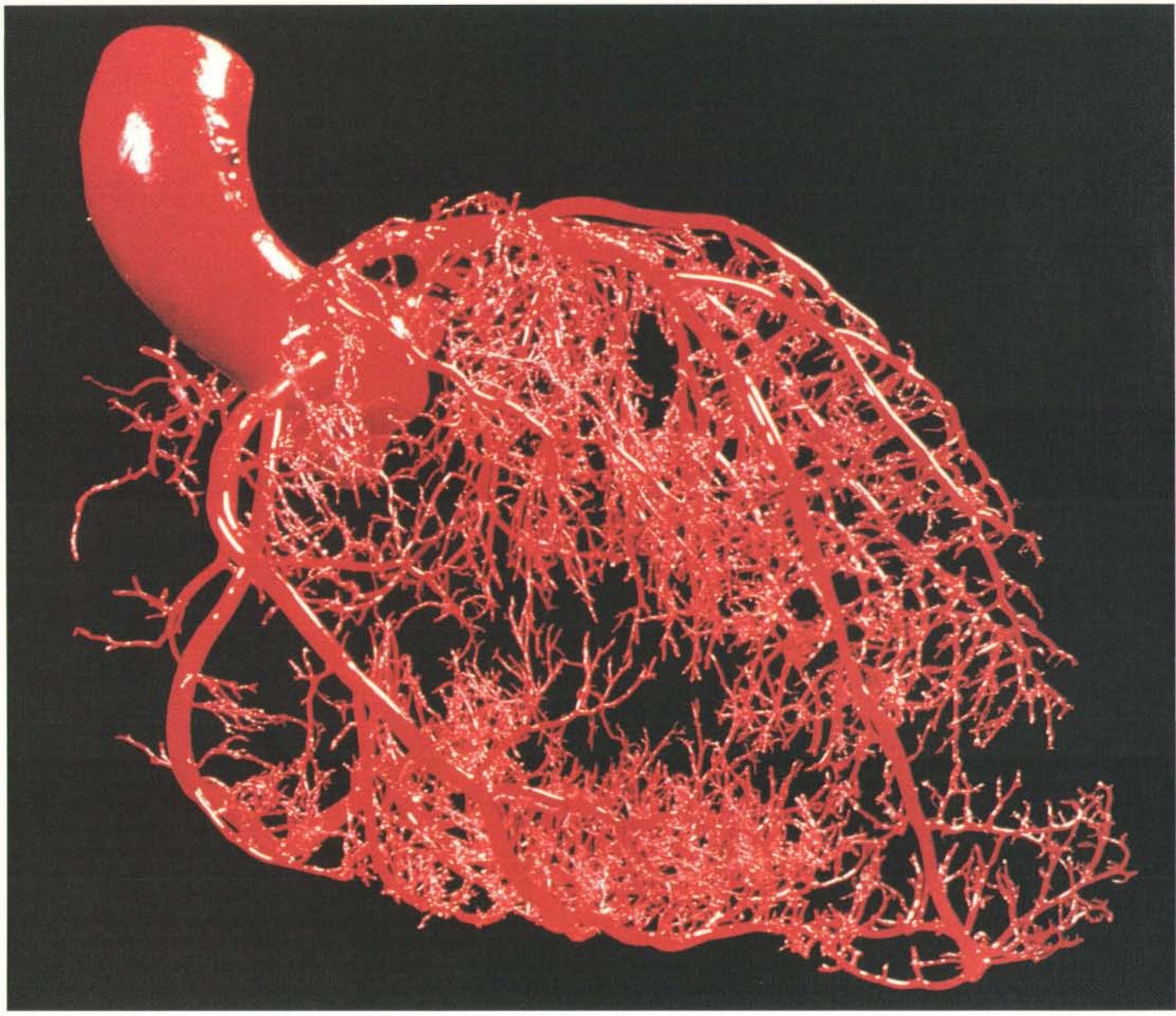


Lines have a dimension of one, circles have two dimensions and spheres have three. But fractals have noninteger (fractional) dimensions. Whereas a smooth Euclidean line precisely fills a one-dimensional space, a fractal line spills over into a two-dimensional space. A fractal line—a coastline, for example—therefore has a dimension between one and two. Likewise a fractal surface—a moutain, for instance—has a dimension between two and three. The greater the dimension of a fractal, the greater the chance that a given region of space contains a piece of that fractal.

In the human body fractallike structures abound in networks of blood vessels, nerves and ducts. The most carefully studied fractal in the body is the system of tubes that transport gas to and from the lungs. In 1962 Ewald R. Weibel and Domingo M. Gomez and later Otto G. Raabe and his co-workers made detailed measurements of the lengths and diameters of tubes in this irregular network of airways. Recently two of us (West and Goldberger) in collaboration with Valmik Bhargava and Thomas R. Nelson of the University of California at San Diego reanalyzed these measurements from the lung casts of humans and several other mammalian species. We found, despite subtle interspecies differences, the type of scaling predicted for the dimensions of a fractal.

***SELF-SIMILARITY** of a system implies that features of a structure or a process look alike at different scales of length or time. When the structures of the small intestine are observed at several different magnifications* (drawings above), *the resemblance between the larger and smaller details suggests self-similarity. When the heart rate of a healthy individual is recorded for three, 30 and 300 minutes* (curves below), *the quick, erratic fluctuations seem to vary in a similar manner to the slower fluctuations.*

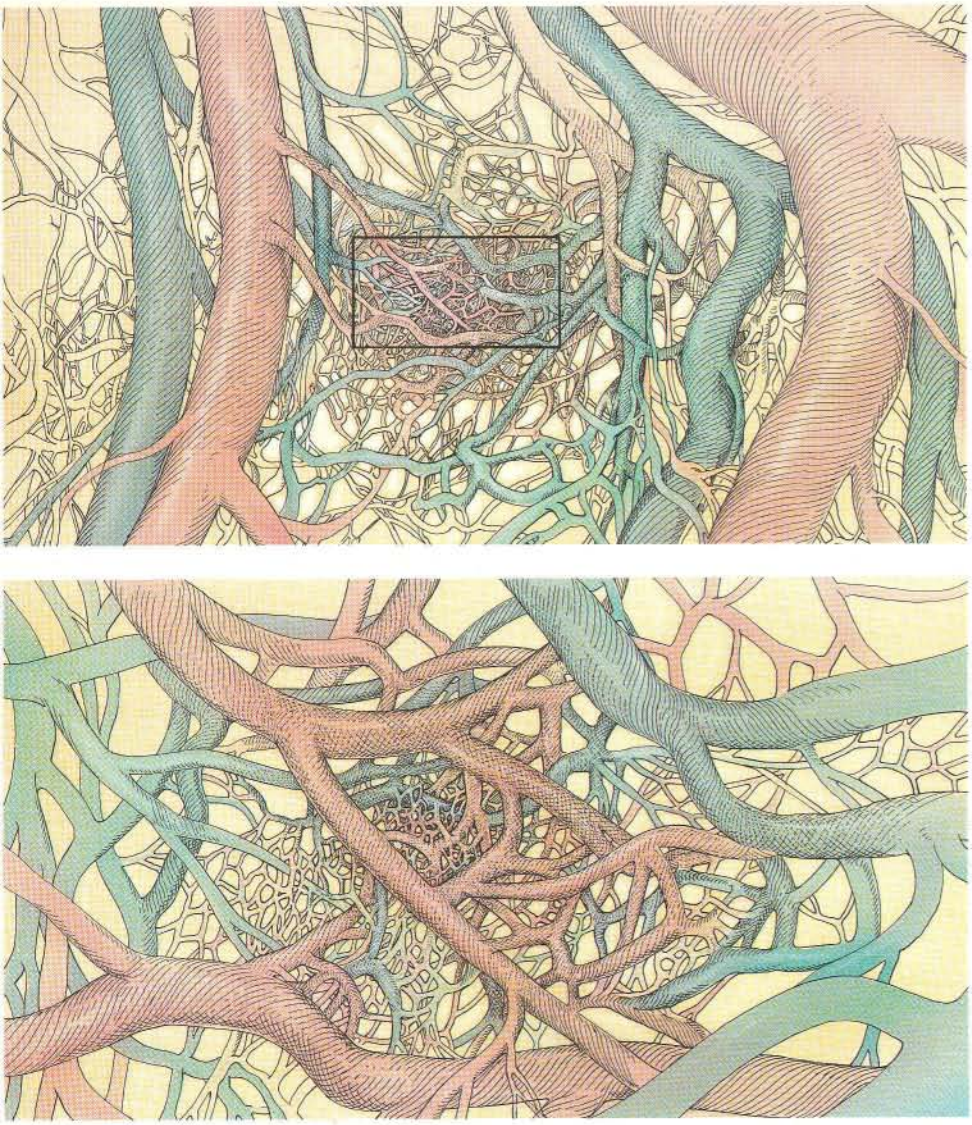




Many other organ systems also appear to be fractal, although their dimensions have not yet been quantified. Fractallike structures play a vital role in the healthy mechanical and electrical dynamics of the heart. First, for example, a fractallike network of coronary arteries and veins conveys blood to and from the heart muscles. Hans van Beek and James B. Bassingthwaighte of the University of Washington recently used fractal geometry to explain anomalies in the blood-flow patterns to the healthy heart. Interruption of this arterial flow may cause a myocardial infarction (heart attack). Second, a fractallike canopy of connective-tissue fibers within the heart—the chordae tendineae—tethers the mitral and tricuspid valves to the underlying muscles. If these tissues break, there can be severe regurgitation of blood from the ventricles to the atria, followed by congestive heart failure. Last, fractal architecture is also evident in the branching pattern of certain cardiac muscles as well as in the His-Purkinje system, which conducts electrical signals from the atria to the cardiac muscles of the ventricles.

Although these fractal anatomies serve apparently disparate functions in different organ systems, several common anatomical and physiological themes emerge. Fractal branches or folds greatly amplify the surface area available for absorption (as in the intestine), distribution or collection (by the blood vessels, bile ducts and bronchial tubes) and information processing (by the nerves). Fractal structures, partly by virtue of their redundancy and irregularity, are robust and resistant to injury. The heart, for example, may continue to pump with relatively minimal mechanical dysfunction despite extensive damage to the His-Purkinje system, which conducts cardiac electrical impulses.

Fractal structures in the human body arise from the slow dynamics of embryonic development and evolution. We have suggested that these processes—like others that produce fractal structures—exhibit deterministic chaos. Recent investigations in physiology have uncovered other examples of apparently chaotic dynamics on shorter, experimentally accessible time scales. In the early 1980's, when investigators began to apply chaos theory to physiological systems, they expected that chaos would be most apparent in diseased or aging systems. Indeed, intuition and medical tradition gave them good reason to think



BLOOD VESSELS of the heart exhibit fractallike branching. The large vessels (cast at left) *branch into smaller vessels* (top drawing), *which in turn branch into even smaller vessels* (bottom drawing).

so. If one listens to the heart through a stethoscope or feels the pulse at the wrist, the rhythm of the heart seems to be regular. For an individual at rest the pulse strength and the interval between heartbeats seem roughly constant. For this reason cardiologists routinely describe the normal heart rate as regular sinus rhythm.

More careful analysis reveals that healthy individuals have heart rates that fluctuate considerably even at rest. In healthy, young adults the heart rate, which averages about 60 beats per minute, may change as much as 20 beats per minute every few heartbeats. In the course of a day the heart rate may vary from 40 to 180 beats per minute.

For at least five decades physicians have interpreted fluctuations in heart rate in terms of the principle of homeostasis: physiological systems normally operate to reduce variability and to maintain a constancy of internal function. According to this theory, developed by Walter B. Cannon of Harvard Medical School, any physiological variable, including heart rate, should return to its "normal" steady state after it has been perturbed. The principle of homeostasis suggests that variations of the heart rate are merely transient responses to a fluctuating environment. One might reasonably postulate that during disease or aging the body is less able to maintain a constant heart rate at rest, so that the magnitude of the variations in heart rate is greater.

A different picture develops when one carefully measures the normal beat-to-beat variations in heart rate and plots them throughout a day. This time-series plot appears ragged, irregular and, at first glance, completely random. But a pattern emerges from the heart-rate data plotted over several different time scales. If one concentrates on a few hours of the time series, one finds more rapid fluctuations whose range and sequence look somewhat like the original, longer time-series plot. At even shorter time scales (minutes), one finds even more rapid fluctuations that again appear to be similar to the original plot. The beat-to-beat fluctuations on different time scales appear to be self-similar, just like the branches of a geometric fractal. This finding suggests that the mechanism that controls heart rate may be intrinsically chaotic. In other words, the heart rate may fluctuate considerably even in the absence of fluctuating external stimuli rather than relaxing to a homeostatic, steady state.

To investigate whether beat-to-beat heart-rate variations are indeed chaotic or periodic, one can compute the Fourier spectrum of the time-series plot for heart rate. The Fourier spectrum of any waveform (such as the time-series plot) reveals the presence of periodic components. If a time-series plot showed a heartbeat of exactly one beat per second, the spectrum would show a sharp spike at a frequency of one beat per second. On the other hand, the time-series plot of a chaotic heartbeat would generate a spectrum that showed either broad peaks or no well-defined peaks. Spectral analysis of normal heart-rate variability in fact shows a broad spectrum suggestive of chaos.

Another tool for analyzing the dynamics of a complex nonlinear system is a "phase space" representation. This technique tracks the values of independent variables that change with time. The number and type of independent variables depend on the system [see "Chaos," by James P. Crutchfield, J. Doyne Farmer, Norman H. Packard and Robert S. Shaw; SCIENTIFIC AMERICAN, December, 1986]. For many complex systems all of the independent variables cannot be readily identified or measured. For such systems phase-space representations can be plotted using the method of delay maps. For the simplest delay map, each point on the graph corresponds to the value of some variable at a given time plotted against the value of that same variable after a fixed time delay. A series of these points at successive times outlines a curve, or trajectory, that describes the system's evolution.

To identify the type of system dynamics (chaotic or periodic), one determines the trajectories for many different initial conditions. Then one searches for an at-

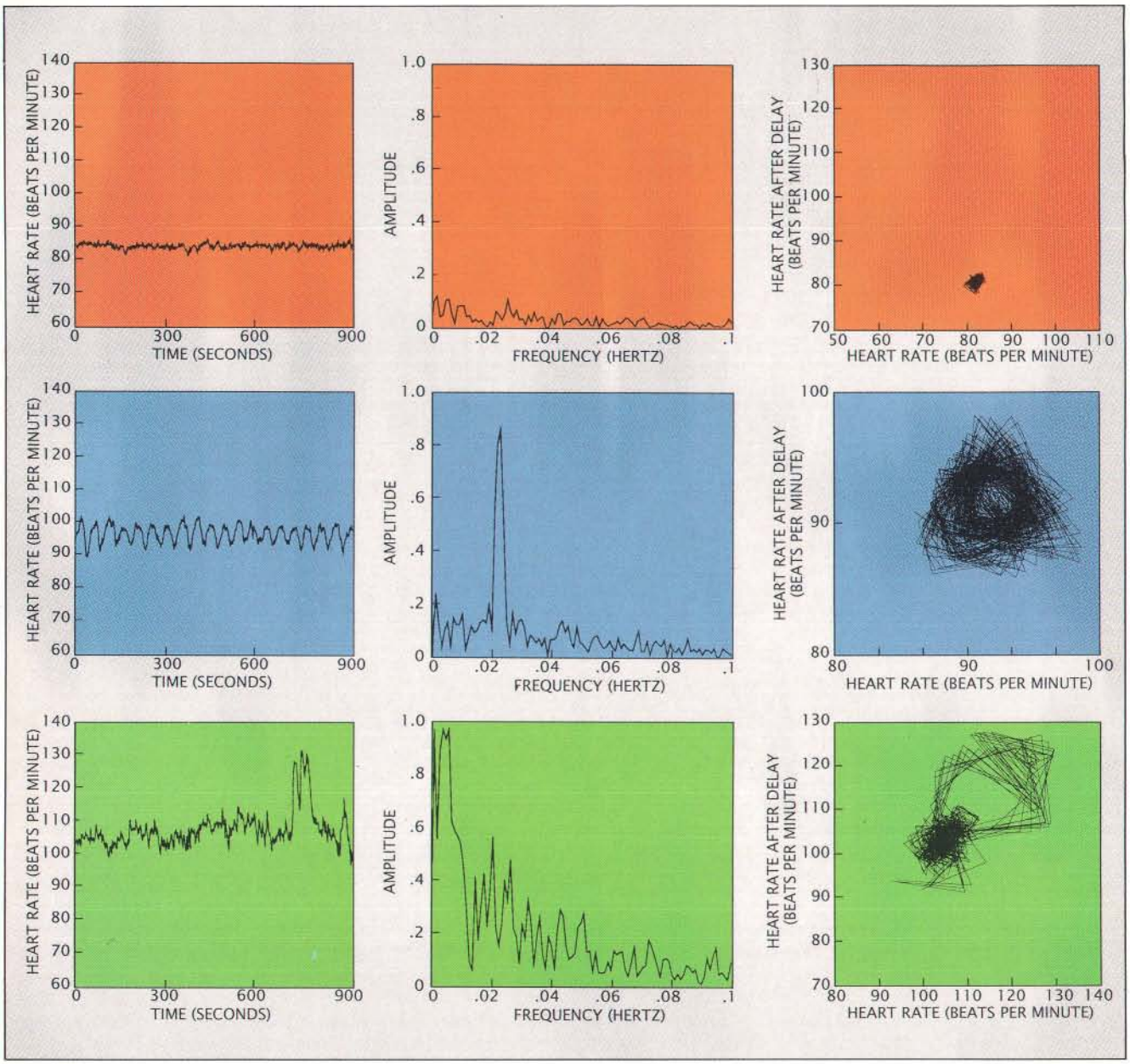
tractor: a region of phase space that attracts trajectories. The simplest kind of attractor is the fixed point. It describes a system—such as a damped pendulum—that always evolves to a single state. In the phase space near a fixed-point attractor, all the trajectories converge to a single point.

The next most complicated attractor is the limit cycle. It corresponds to a system—such as an ideal, frictionless pendulum—that evolves to a periodic state. In the phase space near a limit cycle, the trajectories follow a regular path, for example, one that is circular or elliptical.

Other attractors are simply called "strange." They describe systems that are neither static nor periodic. In the phase space near a strange attractor, two trajectories that started under almost identical conditions will diverge over the short term and become very different over the long term. The system described by a strange attractor is chaotic.

We recently analyzed the phase-space representations for the normal heartbeat. What we found was more like a strange attractor than like the periodic attractor characteristic of a truly regular process. This observation was another indication that the dynamics of the normal heartbeat may be chaotic.

The mechanism for chaos in the beat-to-beat variability of the healthy heart probably arises from the nervous system. The sinus node (the heart's natural pacemaker) receives signals from the involuntary (autonomic) portion of the nervous system. The autonomic nervous system in turn has two major branches: the parasympathetic and the sympathetic. Parasympathetic stimulation decreases the firing rate of sinus-node cells,



HEART RATE is shown as time-series plots **(left),** ***Fourier spectra*** **(center)** ***and phase-space plots*** **(right).** ***A heart rate 13 hours before cardiac arrest*** **(top)** ***is nearly constant as indicated by the flat spectrum and the phase-space trajectory suggestive of a point attractor. A heart rate eight days before sudden cardiac death*** **(middle)** ***is quite periodic as shown by the spike and the trajectory suggestive of a noisy limit cycle. A healthy heart rate*** **(bottom)** ***appears erratic; it has a broad spectrum and a trajectory resembling a strange attractor.***

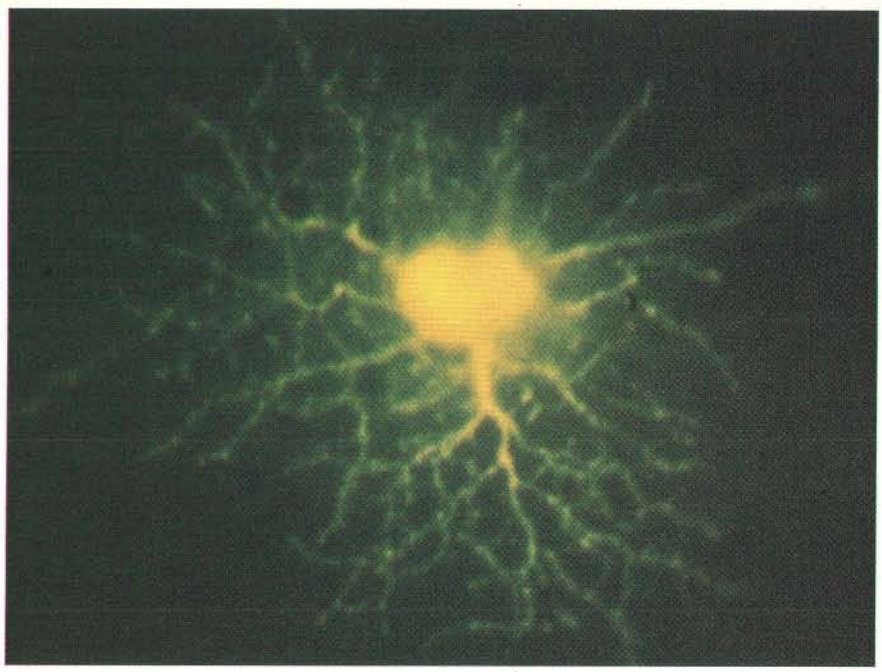
whereas sympathetic stimulation has the opposite effect. The influence of these two branches results in a constant tug-of-war on the pacemaker. The result of this continuous buffeting is fluctuations in the heart rate of healthy subjects. Recently investigators, including Richard J. Cohen and his colleagues at the Massachusetts Institute of Technology, have quantified the reduction in heartbeat variability that occurs after heart transplantation, a procedure in which the autonomic nerve fibers are cut.

Recent evidence from several laboratories suggests that chaos is a normal feature of other components of the nervous system. Gottfried Mayer-Kress of the Los Alamos National Laboratory, Paul E. Rapp of the Medical College of Pennsylvania and Agnes Babloyantz and Alain Destexhe of the Free University of Brussels have analyzed electroencephalograms of healthy individuals and have found evidence for chaos in the nervous system. Otto E. Rössler and his colleagues at the University of Tübingen in West Germany have also discovered indications of chaos in components of the nervous system that are responsible for hormone secretion. They have analyzed temporal changes in hormone levels in healthy human subjects and have found apparently chaotic fluctuations.

Other workers have recently simulated interactions among nerve cells to show how chaos might arise. Walter J. Freeman of the University of California at Berkeley has demonstrated that chaos can be generated in a model of the olfactory system. The model incorporates a feedback loop among the "neurons" and a delay in response times. Earlier, Léon Glass and Michael C. Mackey of McGill University had recognized the importance of time delays in producing chaos.

Why should the heart rate and other systems controlled by the nervous system exhibit chaotic dynamics? Such dynamics offer many functional advantages. Chaotic systems operate under a wide range of conditions and are therefore adaptable and flexible. This plasticity allows systems to cope with the exigencies of an unpredictable and changing environment.

Many pathologies exhibit increasingly periodic behavior and a loss of variability. Early indications that even the dying heart may behave periodically came from Fourier analysis of electrocardiographic waveforms during ventricular tachycardia or ventricular fibrillation, the very rapid cardiac rhythms that most commonly cause cardiac arrest. In the mid-1980's Raymond E. Ideker and his colleagues at the Duke University School of Medicine recorded the waveforms associated with ventricular fibrillation from the innermost layers of the dog heart. They found that the fibrillatory activity inside the heart was a much more periodic process than previously thought.



NEURON exemplifies fractal structure. The cell body branches into dendrites, which in turn branch into finer fibers. This structure may be related to chaos in the nervous system.

In 1988 two of us (Goldberger and Rigney) did a retrospective study of the ambulatory electrocardiograms of people who had severe heart disease. We discovered that the pattern of heartbeats of those patients often became less variable than normal anywhere from minutes to months before sudden cardiac death. In some cases the overall beat-to-beat variability was reduced; in others highly periodic heart-rate oscillations appeared and then stopped abruptly.

Somewhat similarly, the nervous system may show the loss of variability and the appearance of pathological periodicities in disorders such as epilepsy, Parkinson's disease and manic depression. And whereas under normal conditions white-blood-cell counts in healthy subjects have been reported to fluctuate chaotically from day to day, in certain cases of leukemia the white-cell count oscillates periodically.

The periodic patterns in disease and the apparently chaotic behavior in health do not imply that all pathologies are associated with increased regularity. In some cardiac arrhythmias the pulse rate is so erratic that the individual may complain of "palpitations." Some of these events actually represent oscillations that seem irregular but are actually periodic when carefully analyzed. In other arrhythmias the heartbeat is in fact unpredictably erratic. None of these irregular pathologies, however, has been shown to represent nonlinear chaos—although the pulse may feel quite "chaotic" in the colloquial sense.

Physiology may prove to be one of the richest laboratories for the study of fractals and chaos as well as other types of nonlinear dynamics. Physiologists need to develop a better understanding of how developmental processes lead to the construction of fractal architectures and how dynamic processes in the body generate apparent chaos. In the near future, studies of fractals and chaos in physiology may provide more sensitive ways to characterize dysfunction resulting from aging, disease and drug toxicity.

FURTHER READING

AN ESSAY ON THE IMPORTANCE OF BEING NONLINEAR. B. J. West in *Lecture Notes in Biomathematics 62*. Edited by S. Levine. Springer-Verlag, 1985.

FRACTALS IN PHYSIOLOGY AND MEDICINE. Ary L. Goldberger and Bruce J. West in *Yale Journal of Biology and Medicine*, Vol. 60, pages 421–435; 1987.

PHYSIOLOGY IN FRACTAL DIMENSIONS. Bruce J. West and Ary L. Goldberger in *American Scientist*, Vol. 75, No. 4, pages 354–365; July–August, 1987.

NONLINEAR DYNAMICS IN SUDDEN CARDIAC DEATH SYNDROME: HEARTRATE OSCILLATIONS AND BIFURCATIONS. A. L. Goldberger, D. R. Rigney, J. Mietus, E. M. Antman and S. Greenwald in *Experientia*, Vol. 44, pages 983–987; 1988.

How Plants Make Oxygen

A biochemical mechanism called the water-oxidizing clock enables plants and some bacteria to exploit solar energy to split water molecules into oxygen gas, protons and electrons

by Govindjee and William J. Coleman

Because molecular oxygen is a necessity of life for human beings, it is easy to forget that simple organisms lived without it for hundreds of millions of years. For those early anaerobic organisms, oxygen was a toxic substance that could steal essential electrons from the molecules in their cells. It may seem surprising, then, that many of these anaerobic cells engaged in a form of photosynthesis, because photosynthesis produces all the oxygen in the atmosphere. The exact process by which oxygen is generated in photosynthesis has been something of a mystery, but now the mechanism can be described in some detail. It is a "water-oxidizing clock" that generates a molecule of oxygen with every four ticks.

The fundamental task of photosynthesis is to make it possible for cells to convert carbon dioxide into carbohydrates with energy absorbed from the sun. The production of oxygen is not crucial, which is why anaerobic cells could photosynthesize without making molecular oxygen long ago and why they have continued to do so to this day.

GOVINDJEE and WILLIAM J. COLEMAN have collaborated on studies of the mechanism of oxygen evolution in green plants. Govindjee is professor of biophysics and plant biology at the University of Illinois at Urbana-Champaign. He was born in Allahabad, India, and received his Ph.D. in biophysics from Urbana in 1960. He has written many books, papers and reviews on the subject of photosynthesis, including two previous articles for *Scientific American*. Coleman is a National Science Foundation postdoctoral fellow studying bacterial photosynthesis in the department of chemistry at the Massachusetts Institute of Technology. He received his bachelor's degree in biology and literature from the University of Pennsylvania in 1979 and his Ph.D. from Urbana in 1987 for his work with Govindjee.

If oxygen is toxic, why and how did green plants and their ancestors ever begin producing it through photosynthesis? The answer to the first question involves energy metabolism. Sunlight provides the energy that drives life on the earth, but cells cannot store or employ that light energy directly; it must be converted into a more usable, chemical form. Electrons are part of the common "currency" of biological energy conversion: many energetic reactions in cells can be generally understood as the transfer of electrons between molecules.

To live, therefore, cells need a source of electrons. Anoxygenic photosynthetic bacteria typically oxidize, or draw electrons from, organic acids and simple inorganic compounds. These substances are relatively rare, however; consequently, anoxygenic bacteria survive today only in sulfur springs, lake bottoms and similar environments where such molecules are sufficiently plentiful.

Approximately three billion years ago, however, some photosynthetic cells learned how to spread into virtually any environment by tapping the electrons in a nearly ubiquitous substance: water. They evolved the ability to split pairs of water (H_2O) molecules into electrons, protons (H^+'s, or hydrogen nuclei) and molecular oxygen (O_2). The electrons and protons were energetically useful; the O_2 was simply a by-product. In short, the evolution of O_2 was a breakthrough for photosynthetic organisms not because O_2 was important in itself but because it meant that photosynthetic cells could exploit water and invade new, more diverse environments.

How cells make oxygen is a far more complicated question. Developing the ability to draw on water as an electron source was no simple feat, and it demanded several modifications to the established photosynthetic mechanism. Because water molecules give up their electrons only grudgingly, the relatively weak oxidant (electron-accepting molecule) that anoxygenic photosynthetic bacteria were able to generate with sunlight had to be replaced with a much stronger one. Even so, the energy available from a single photon, or quantum unit, of visible light is not sufficient to split a water molecule. This problem appears to have been solved by drawing the energy from four photons to split two water molecules, thereby releasing four electrons and four protons. Such a mechanism creates yet another difficulty, however, because the photochemical apparatus can handle only one electron at a time.

To solve that problem, photosynthetic cells developed the special water-splitting catalyst that we call the water-oxidizing clock: a unique biochemical ratcheting mechanism for stabilizing intermediate stages of the water-splitting reaction so that the electrons could be transferred one by one. In recent years much has been learned about the workings of the water-oxidizing clock and its place in the overall process of photosynthesis.

In higher plants, the primary reactions of photosynthesis take place within specialized thylakoid membranes inside the cell structures called chloroplasts. Embedded in the thylakoid membranes are various protein complexes, each of which contributes to the total photosynthetic reaction. The generation of O_2 takes place entirely within the complex of proteins

OXYGEN-RICH BUBBLES on the leaves of a submerged green plant show that photosynthesis is taking place. The oxygen (O_2) is a product of a light-driven reaction in which pairs of water (H_2O) molecules are stripped of four electrons (e^-) and four protons (H^+). Anoxygenic photosynthetic bacteria cannot split water molecules in this way and must get essential electrons from other sources.

and pigments known as photosystem II, which is found in the cells of all oxygenic photosynthesizers: cyanobacteria, algae and other plants containing chlorophyll pigments.

The essential task of photosystem II is to act as a tiny capacitor, storing energy by separating and stabilizing positive and negative charges on either side of the thylakoid membrane. To do this, an array of specialized pigments in photosystem II absorbs a photon and efficiently converts this light energy into a widening separation of charge.

Orchestrating the movements in the complex process of converting light energy into a separation of charge requires the collaboration of many specialized polypeptides and proteins in the photosystem. Polypeptides are linear polymers of amino acids arranged in a defined sequence; they are often many hundreds of amino acids in length. Proteins consist of one or more polypeptides folded into intricate, orderly structures.

The electron-transfer reactions in photosystem II take place within the so-called reaction center. The major structural components of the reaction center are the large polypeptides named D1 and D2 and a smaller protein named cytochrome b_{559}. A polypeptide with a molecular weight of 33 kilodaltons and at least two others of different weights are bound to the inner surface of the thylakoid membrane. These polypeptides serve as a stabilizing matrix for the pigments and other molecules in photosystem II that perform the electron-transfer and oxygen-producing reactions. Other small polypeptides are associated with photosystem II, but their functions are still unknown. Several organic ions and charged atoms—manganese, chloride, calcium, iron and bicarbonate—are involved in catalyzing electron transfer, maintaining the protein structure or regulating the photosystem's activity.

In addition, large numbers of "antenna" chlorophyll molecules collect light energy and funnel it efficiently to the reaction center. Several hundred antenna pigment molecules are associated with each reaction center.

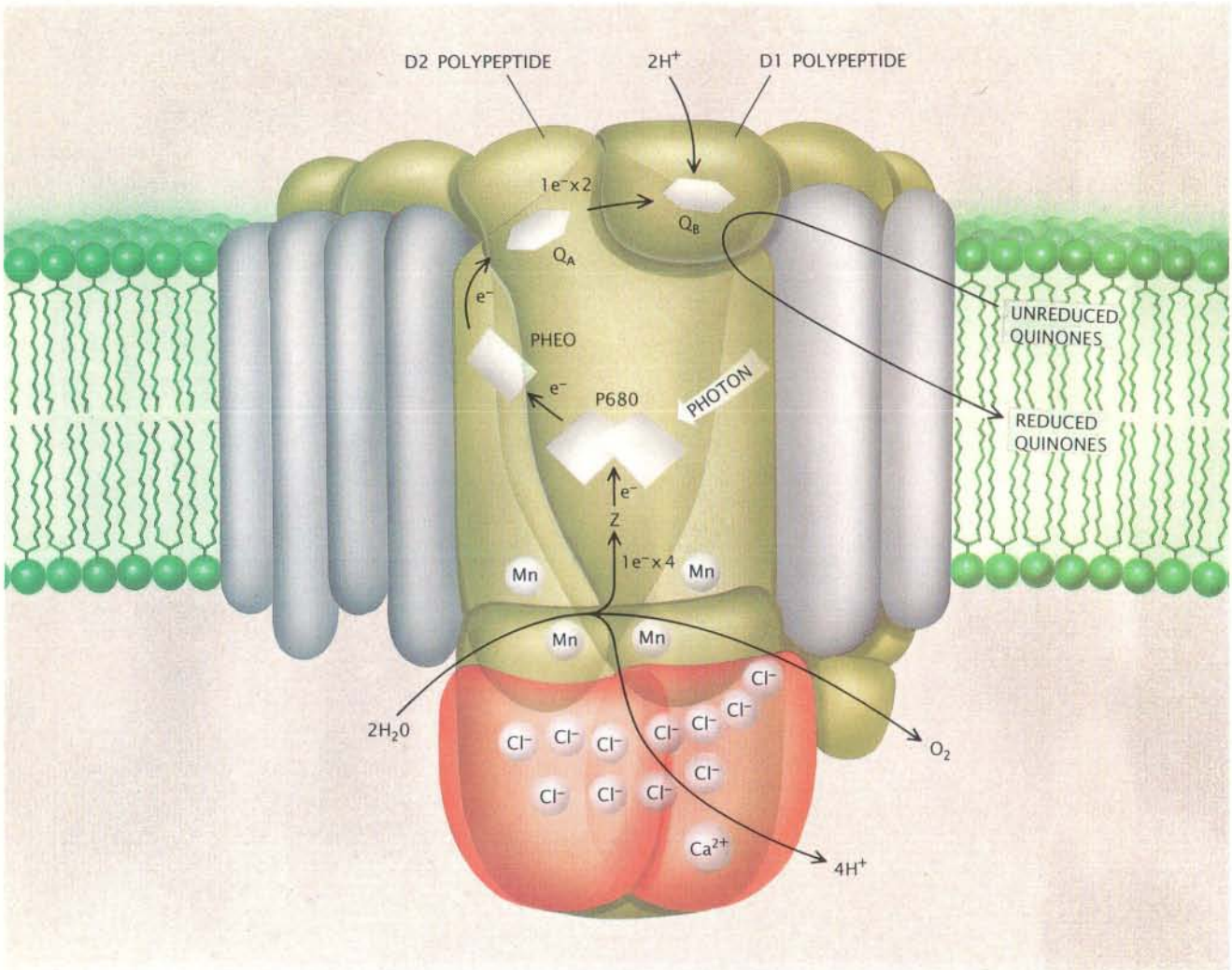
Because the structure of photosystem II is so complicated, many advances in understanding it have come from studies of equivalent complexes in photosynthetic bacteria. The work of Johann Deisenhofer, Robert Huber and Hartmut Michel, who determined the structure of the photosynthetic reaction center in the bacterium *Rhodopseudomonas viridis*, earned them the Nobel prize for chemistry in 1988.

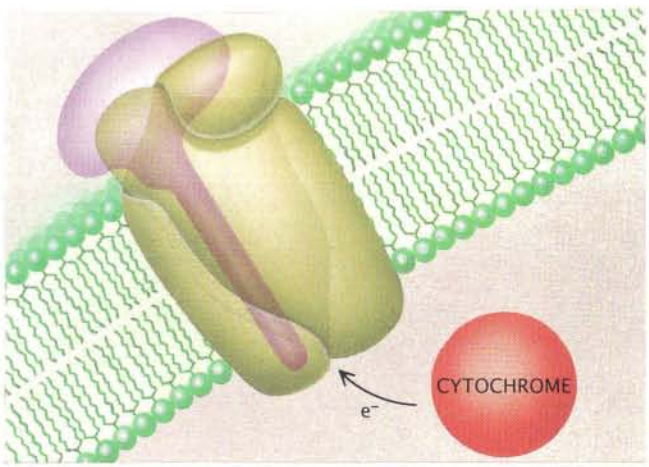
There are many differences between bacterial and green-plant photosynthetic complexes. As previously noted, bacteria do not produce molecular oxygen; moreover, they depend not on chlorophyll but on the pigment bacteriochlorophyll, which absorbs light maximally at a much longer wavelength and is a much weaker oxidant. Yet bacterial photosynthetic complex-

es do catalyze the essential reaction that converts light into an electrochemical potential gradient across a biological membrane [see "Molecular Mechanisms of Photosynthesis," by Douglas C. Youvan and Barry L. Marrs; SCIENTIFIC AMERICAN, June, 1987].

From these bacterial studies, it appears that the electron-transporting mechanism in the reaction center of photosystem II has five components: a chlorophyll pigment that acts as the primary donor of an electron; a secondary electron donor named Z that reduces the chlorophyll (that is, it replaces the electron that the chlorophyll has lost); pheophytin, a pigment that accepts an electron from the chlorophyll; a primary plastoquinone electron acceptor, Q_A; and a secondary quinone electron acceptor, Q_B.

The chlorophyll pigment in the reaction center is believed to consist of a special pair of chlorophyll molecules that seem to be chemically identical to many of the antenna pigments but are functionally different. The pigment is named P680 because it most strongly absorbs light that has a wavelength of 680 nanometers.





PHOTOSYSTEM II is the protein-pigment complex in the thylakoid membrane of chloroplasts that produces oxygen and traps light energy. The flow of electrons through the photosystem (*above*) is driven by the absorption of a photon by a special pair of chlorophylls (P680). The process involves many other molecules, including pheophytin (Pheo) and two kinds of quinone (Q_A, Q_B) molecules, and such metal atoms and ions as manganese (Mn), chloride (Cl^-) and calcium (Ca^{2+}). In this simplified model, several pigment molecules have been omitted for clarity. Photosystem II reactions near the inner side of the membrane take four electrons from two water molecules and make molecular oxygen; these reactions constitute the water-oxidizing clock. Protons released from these reactions contribute to the synthesis of adenosine triphosphate. A simpler photosystem from an anoxygenic bacterium (*right*) lacks a water-oxidizing clock. Electrons from compounds other than water are passed to this photosystem by a cytochrome protein.

In 1988 the work of Bridgette A. Barry and Richard J. Debus, who were then at Michigan State University, and Willem F. Vermaas of Arizona State University and their colleagues helped to identify Z as one of the amino acids (known as a tyrosine) within the D1 polypeptide. The quinone Q_A is tightly bound to the photosystem II complex, but Q_B can diffuse freely between protein complexes in the membrane when it has accepted two electrons.

During photosynthesis the antenna pigments absorb a photon and funnel this energy to the P680 in the reaction center. There, the excitation energy is converted into a charge separation when the P680 enters an excited state and quickly passes one electron to a nearby pheophytin molecule [*see illustration on next page*]. The pheophytin now carries an excess negative charge, whereas there is a positively charged "hole" on the P680 because it has lost an electron; the P680 has become a $P680^+$. The separation of the charges gets wider when the pheophytin passes its extra electron to Q_A. The distance increases still more when Z donates an electron to the $P680^+$ and picks up the positive charge and the Q_A donates its extra electron to Q_B.

The transfers of charge take place rapidly, especially the initial transfer of an electron from the excited P680 to pheophytin, which occurs within a few trillionths of a second. This was shown by one of us (Govindjee) in collaboration with Michael R. Wasielewski and Douglas G. Johnson of the Argonne National Laboratory and Michael Seibert of the Solar Energy Research Institute in Golden, Colo.

The stepwise transfer of electrons succeeds in pulling far apart the mutually attractive positive and negative charges. Yet the photosynthetic cycle in photosystem II is not complete until all the components of the reaction mechanism are electrically neutral again and ready to begin the charge-separation process anew. How does the Q_B eliminate its negative charge, and how does the Z regain the electron it has lost?

At the Q_B end of the system, the answer is relatively simple. After the Q_B has acquired two electrons and two protons through two photon-absorption cycles, the doubly reduced Q_B diffuses out of the photosystem II complex and is replaced by an unreduced Q_B. The electrons and protons on the freely moving Q_B are carried to still another complex in the photosynthetic pathway. The protons released on the inner side of the thylakoid membrane are eventually exploited to make adenosine triphosphate, an energy-storing compound essential for cellular metabolism.

At the opposite end of photosystem II, it is much harder for the Z to obtain the electron it needs to return to its original state. The electron must come from some oxidizable substance that is available in the cell's environment. Organic acids (such as acetate, malate and succinate) and simple inorganic compounds (such as sulfide and thiosulfate) can be good electron sources, and they are the ones that anoxygenic photosynthetic bacteria exploit; in these bacteria, which lack Z, a cytochrome protein shuttles an electron to the oxidized special pair of chlorophylls in the reaction center.

A molecule far more abundant than organic acids are—and therefore a potentially richer source of electrons—is water. Yet, although the oxidizing strength of $P680^+$ is great, it is not sufficient by itself to strip water molecules of their electrons.

The problem is that the water-oxidizing reaction releases four electrons simultaneously, whereas $P680^+$ can accept only one electron at a time. It therefore became clear to investigators a few decades ago that there must be a catalytic site near Z and the P680 that can, in effect, prolong the oxidation reaction. This water-splitting catalyst must associate with pairs of water molecules and stabilize them during a gradual oxidation process in which electrons are taken away one at a time. The search for this mechanism eventually led to the discovery of the water-oxidizing clock.

An important clue to how this mechanism works was derived from the observation that not all the electrons reach the $P680^+$ chlorophyll molecule at the same rate. Instead the observed transfer time for the electrons varies periodically. This has been demonstrated by experiments in which membranes containing photosystem II reaction centers are placed in darkness and then exposed to brief flashes of light. Each flash is not only very intense but also as brief as possible, so that it sends (on average) only one photon into the photosystem. The observed result is that $P680^+$ recovers an electron in darkness at different rates, depending on the number of light flashes.

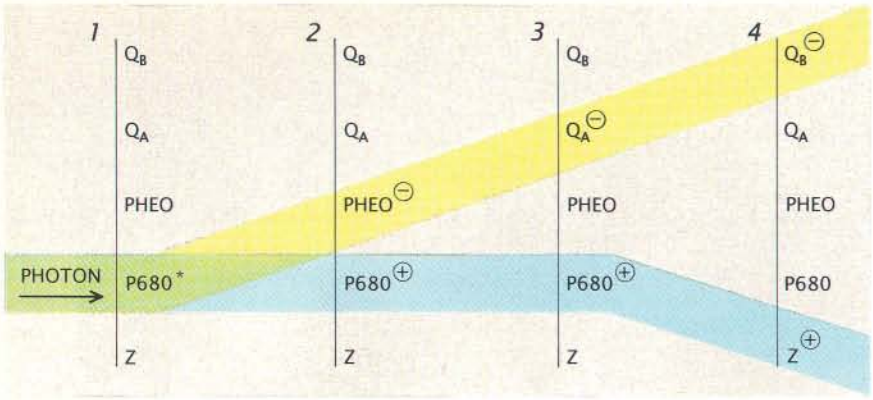
For example, the time for half of the $P680^+$'s to convert back to P680's is approximately 20 billionths of a second after the first and fifth flashes but is much longer after the second, third and fourth flashes. The change in recovery time varies cyclically every four flashes. The four-point periodicity suggests that a cyclic reaction with four steps donates electrons to the reaction center.

These studies of the behavior of P680 were especially important. Pierre Joliot of the Institute of Physicochemical Biology in Paris had previously demonstrated in 1969 that there was a four-point periodicity in the photosynthetic production of oxygen. With a highly sensitive platinum-electrode measuring device that responded to traces of oxygen, Joliot measured the amount of gas that evolved after a series of flashes. There was no O_2 evolution after the first flash and none (or very little) after the second, but there was a maximum release after the third. Thereafter, the amplitude of the O_2 yield oscillated with a period of four until the differences gradually were damped out, or diminished.

Bessel Kok of the Martin Marietta Laboratories in Baltimore proposed a simple hypothesis in 1970 to explain Joliot's results, an idea that came to be known as the water-oxidizing clock or cycle. Kok suggested that the oxygen-producing complex in photosystem II could exist in several different, transient states of oxidation that he called S states. He could not define the chemical nature of the S states precisely, but he hypothesized that each S state made a unique contribution to a four-stage cyclic mechanism.

Kok suggested that, in darkness, the clock settles into one of two S states: S_0 or S_1. The predominant (and more stable) state is S_1, which has one more oxidizing equivalent than S_0; in other words, the complex of molecules corresponding to S_1 has one fewer electron than does the S_0 complex. The chemical basis for the predominance of S_1 is not known.

After one flash of light, P680 becomes $P680^+$ and must eventually be reduced by an electron. Kok hypothesized that the clock must undergo a change that boosts it into the next-highest oxidation state: a clock that starts at S_1 goes to S_2, and one that starts at S_0 goes to S_1. The transition occurs because one electron is released from the clock to convert $P680^+$ back to P680. A second flash creates another $P680^+$ and boosts the S_2's to S_3's, and a third flash converts the S_3's to S_4's. When the clock reaches the S_4 state, it has released four electrons and is ready to complete the water-



STEPWISE ELECTRON TRANSFER in the photosystem II reaction center stores some light energy in the form of separated positive and negative charges. When the P680 absorbs a photon, it enters an excited state and becomes P680* (*1*). The P680* donates an electron to a pheophytin molecule and is left with an electron deficit, or positive charge (*2*). The pheophytin's negative charge is passed to a primary quinone molecule, Q_A (*3*). Finally, the P680+ takes an electron from Z, an amino acid, and Q_A passes its extra electron to Q_B, another quinone molecule (*4*). The electron-transfer chain returns to its original state when Z accepts an electron from the water-oxidizing clock and the doubly reduced Q_B is replaced by an unreduced quinone.

splitting reaction. The clock then removes four electrons from the two bound water molecules, releases O_2 and drops from S_4 back to S_0, making it possible for the cycle to begin again.

This situation is not unlike that of a base runner in baseball: the player must tag all four bases in sequence to end up where he started. If the player misses a tag, he may be forced to retreat; similarly, there is a possibility that the clocks will not progress smoothly from one state to the next. A small probability exists, for example, that S_1 will not change to S_2 after a flash because the photosystem did not utilize the photon efficiently. There is also a low probability that a photosystem will absorb two photons during one flash (if the flashes are not extremely short) and that the water-oxidizing clock may advance in one step from S_1 to S_3 via S_2.

Kok's mechanism explained Joliot's observations of the clock's oxygen-producing behavior. Because most of the clocks in a dark-adapted sample are in the S_1 state, the maximum release of oxygen takes place after the third flash, when the clocks change from S_3 to S_4 to S_0 and spontaneously release oxygen. The clocks that began in the S_0 state release O_2 after the fourth flash, which is why there is a small oxygen release then.

Those random "errors" that occur when a few clocks fail to advance during a flash or when they advance by two S states can account for the gradual damping-out in the O_2-release oscillations. These processes slowly desynchronize the turnover of the clocks in the sample. After many flashes, there is an equilibrium such that the numbers of S_0, S_1, S_2 and S_3 clocks are roughly equal, and the yield of oxygen after each flash remains steady. The situation is analogous to that of a room filled with grandfather clocks: initially they may all chime loudly and synchronously on the hour, but as the clocks variously gain or lose time, the room begins to reverberate with a continuous soft chime.

Joliot and Kok's discovery of the water-oxidizing clock replaced the black box of oxygen production with a new theoretical mechanism. The theory did not explain, however, the physical makeup of the clock or the interactions of the clock with water molecules. A long search soon began for the chemical nature of the charge accumulator in the clock—the material or materials whose variable oxidation states constitute each of the S states.

From the beginning it was assumed that this elusive chemical entity was a metal atom. Protein-bound atoms of transition metals, such as manganese, iron and copper, are good candidates for catalyzing oxidation-reduction reactions because of their ability to donate and accept electrons alternately.

Manganese (Mn) is believed to make up at least part of the charge accumulator because, as has been long known, O_2 production does not take place unless there are four atoms of manganese in photosystem II for every P680 molecule. Manganese is known to catalyze electron-transfer reactions in other enzymes. It can also assume several relatively stable oxidation states, from +2 to +7; that is, manganese ions can variously share between two and seven electrons with other atoms. When the metal is bound to a large molecule such as a protein, these oxidation states are usually abbreviated as Mn(II), Mn(III) and so on.

Metal-containing proteins have been analyzed extensively by the general technique called spectroscopy because some metal complexes absorb particular forms of electromagnetic radiation. If this absorption is measured carefully, it can serve as a spectroscopic "fingerprint" of the protein-bound metal and provide clues to its nuclear or electronic structure. Spectroscopy is especially well suited to the study of manganese compounds. Many of the biologically relevant manganese complexes are "paramagnetic": the manganese atom contains electrons with unpaired spins, and these electrons, like tiny bar magnets, interact strongly with applied external magnetic fields.

Several highly sensitive measuring techniques have exploited the paramagnetic properties of manganese, most notably electron paramagnetic resonance (EPR) spectroscopy. With EPR, changes in the electronic structure of the manganese complex that follow the absorption of light by photosystem II have been studied. Another informative approach has been nuclear magnetic resonance (NMR) spectroscopy, which can measure the properties of the manganese atoms indirectly by monitoring protons in the water molecules that are in contact with the manganese. While working at the University of Illinois at Urbana-Champaign in the mid-1970's, Thomas J. Wydrzynski pioneered the use of NMR to study dynamic changes in the oxidation state of manganese.

X-ray spectroscopy techniques have made valuable contributions to the study of oxidation states and of the physical environment of the manganese atoms in photosystem II. Other studies of the chemical composition of S states have used optical spectroscopy, because manganese complexes have unique absorption bands in the ultraviolet region of the electromagnetic spectrum.

It is worth pointing out, however, that despite the wide range of applicable spectroscopic techniques, two major difficulties have humbled scientists studying the photosynthetic membrane. First, the membrane is complex, and many of its components have overlapping absorption spectra.

Second, because neither the structure nor the chemical nature of the photosystem II complex is known precisely, the data from spectroscopic analyses of the water-oxidizing clock cannot be interpreted definitively. As a result, we do not yet know conclusively what chemically constitutes the various S states. It has been possible, nonetheless, to develop a tentative picture.

It is clear that the manganese atoms undergo dynamic changes, including changes in their oxidation states, during the S-state transitions. A four-point periodicity has been observed in the manganese oxidation-state changes, as Kok's model had suggested. One surprising discovery is that the manganese atoms do not become consistently more oxidized throughout the cycle. S_2 is more oxidized than S_1, and S_1 is more oxidized than S_0, but there is no discernible change in the manganese oxidation states between S_2 and S_3. It seems, then, that the positive charge the clock acquires during its transition from S_2 to S_3 must be carried on some feature of the clock other than the manganese atoms. One of us (Govindjee), together with Subhash Padhye, Takeshi Kambara and David N. Hendrickson of the University of Illinois at Urbana-Champaign, proposed in 1986 that the amino acid histidine in one of the proteins in the clock could possibly store a positive charge.

Work by Melvin P. Klein, Kenneth Sauer and their co-workers at the University of California at Berkeley and by Robert R. Sharp and his colleagues at the University of Michigan at Ann Arbor has helped define the oxidation states of some of the manganese atoms more precisely. Tentatively, S_0 has been identified with the presence of Mn(II), S_1 with Mn(III) and S_2 with Mn(IV). Both Mn(II) and Mn(III) appear to be stable and long-lived in photosystem II; these observations corroborate Kok's prediction of stable S_0 and S_1 states. In contrast, the Mn(IV) associated with S_2 is a relatively transient intermediate. Recent evidence collected in the laboratory of Horst T. Witt of the Technical University in West Berlin indicates that during the S_0-to-S_1 transition, an Mn(II) ion converts to an Mn(III) ion. The only conversions observed during subsequent transitions are from Mn(III) to Mn(IV).

Low-temperature EPR studies by G. Charles Dismukes and Yona Siderer of Princeton University suggest that the S_2 and S_3 states involve multinuclear complexes with as many as four manganese atoms. For example, the S_2 state may be a mixed-valence group consisting of one Mn(III) atom and one Mn(IV) atom or of three Mn(III) atoms and one Mn(IV) atom.

In summary, dynamic changes in the oxidation states of the manganese atoms bound within photosystem II unquestionably correspond to changes in the S states of Kok's clock. The precise chemical and electronic configurations of these states are still uncertain and are under study.

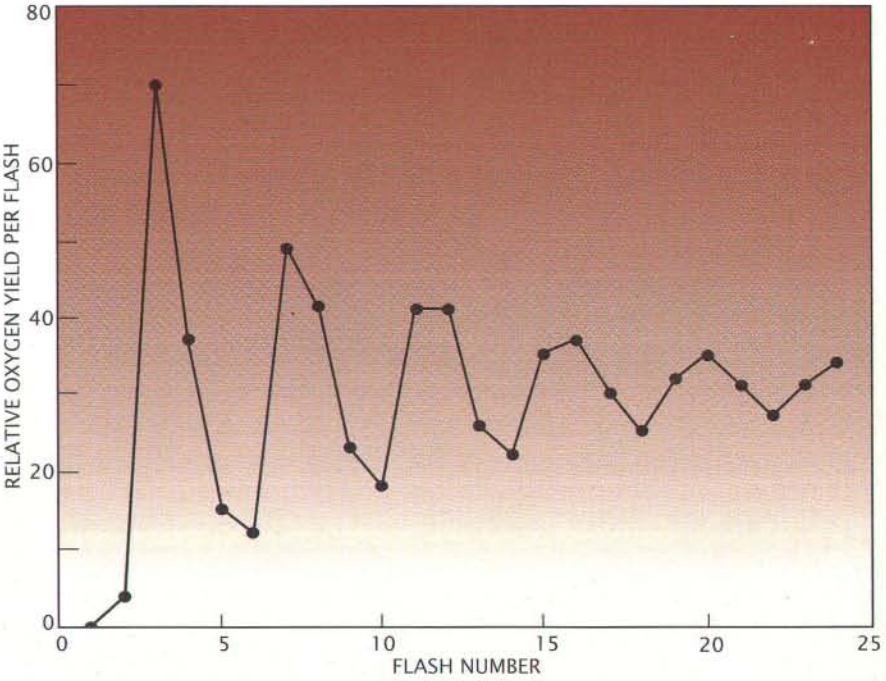
Various experiments have indicated that manganese is probably not bound directly to any of the small polypeptides in the photosystem II complex. This leaves the large D1 and D2 polypeptides as the most likely sites for manganese binding. We have recently proposed that four manganese-binding sites may exist on the D1 and D2 polypeptides on the inner side of the thylakoid membrane, but other investigators have suggested that manganese is bound across the interface between the D1, D2 and 33-kilodalton polypeptides.

X-ray spectroscopy by Klein, Sauer and their collaborators at Berkeley and by Graham N. George and Roger C. Prince of Exxon Research in Annandale, N.J., has revealed some details of the arrangement of the manganese atoms. In the S_1 state, two of the atoms appear to be part of a binuclear complex and are separated by only 2.7 angstroms. (One angstrom is one ten-millionth of a millimeter.) The other pair of manganese atoms is separated by a larger distance. The atoms can be imagined as being at the four corners of a trapezoid.

Because of all these studies, much more is now known about how manganese atoms might catalyze the removal of electrons from water to reduce $P680^+$. Electrons, however, are not the whole story: the water-splitting reaction also produces four protons. Are all four protons released at once, simultaneously with the release of O_2, or are they liberated sequentially along with the electrons?

This question has been answered by careful measurements of proton release in response to a series of flashes. Because the release of protons increases the acidity of the surrounding fluid, the timing of proton release can be studied with electrodes and dyes that are highly sensitive to acidity. C. Frederick Fowler of Martin Marietta and, soon thereafter, Satham Saphon and Anthony R. Crofts, then at the University of Bristol in England, discovered that the four protons are released sequentially. One is released during the S_0-to-S_1 transition, none during S_1-to-S_2, one during S_2-to-S_3 and two during the S_3-to-S_4-to-S_0 transition.

These findings have important im-



YIELD OF OXYGEN from photosynthetic membranes exposed to a series of brief flashes oscillates with a four-point periodicity. It is highest after the third flash and peaks again four flashes later, but the variation in the amplitude gradually decreases as the number of flashes increases. The occurrence of the peaks and the damping of the oscillation are explained by the four-step cycle of the water-oxidizing clock.

STONEHENGE AND THE SPACE TELESCOPE

Over 4,100 years ago a Neolithic people built a remarkable monument on the Salisbury Plain in what is now southern England. As an engineering feat alone, Stonehenge stands as one of the wonders of the world. But a recent discovery has revealed that it served not only as a temple, but as an astronomical computer.

We know very little about the life of the people who built Stonehenge. But one thing that has become increasingly evident is that they were far more sophisticated than was previously believed. Even though they

worked only with Stone Age technology, they built a monument which apparently acted as an astronomical clock. With Stonehenge they could predict eclipses, the exact days of the solstices, the long-term cycles of the moon and sun, and other important heavenly events. They could begin to understand that the universe had order and how it worked.

The need to understand the workings of the universe is very ancient in man. One might even say that it is instinctual, that it is part of what makes us human.

A leap of forty-one centuries and we find ourselves still confronted with the same questions that drove the prehistoric Britons to build Stonehenge. How does the universe work? How did it begin? Will it ever end?

The Hubble Space Telescope will help us solve these primeval mysteries. Once in Earth orbit, the telescope will be able to detect objects as far as fourteen billion light-years away, which is to see fourteen billion years into the past; past the birth of the Earth; past the birth of our galaxy; to the very beginning of time.

The Space Telescope represents a momentous leap in the history of mankind. The builders of Stonehenge must have felt themselves on the verge of the same kind of moment as they discovered that creation actually had order. Within our own grasp is a view of the creation itself.

Lockheed
Giving shape to imagination.

SCIENTIFIC AMERICAN

In Other Languages

LE SCIENZE
L. 3,500/copy L. 35,000/year L. 45,000/(abroad)
Editorial, subscription correspondence:
Le Scienze S.p.A., Via G. De Alessandri, 11
20144 Milano, Italy
Advertising correspondence:
Publietas, S.p.A., Via Cino de Duca, 5,
20122 Milano, Italy

サイエンス
Y950/copy Y10,440/year Y14,000/(abroad)
Editorial, subscription, advertising correspondence:
Nikkei Science, Inc.
No. 9-5, 1-Chome, Otemachi
Chiyoda-ku, Tokyo, Japan

INVESTIGACION Y CIENCIA
500 Ptas/copy 5500 Ptas/year 6200 Ptas (abroad)
Editorial, subscription, advertising correspondence:
Prensa Cientifica S.A.,
Calabria, 235-239
08029 Barcelona, Spain

POUR LA SCIENCE
27FF/copy 265FF/year 315FF/year (abroad)
Editorial, subscription, advertising correspondence:
Pour la Science S.A.R.L.,
8, rue Férou,
75006 Paris, France

Spektrum DER WISSENSCHAFT
9.80 DM/copy 99 DM/year 112.20 DM/(abroad)
Editorial, subscription correspondence:
Spektrum der Wissenschaft GmbH & Co.
Moenchhofstrasse, 15
D-6900 Heidelberg,
Federal Republic of Germany
Advertising correspondence:
Gesellschaft Für Wirtschaftspublizistik
Kasernenstrasse 67
D-4000 Duesseldorf,
Federal Republic of Germany

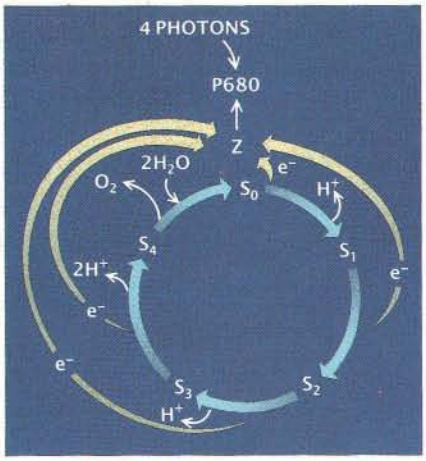
科学
3.80RMB/copy 45.60RMB/year $48/(abroad)
Editorial subscription correspondence:
ISTIC-Chongqing Branch, P.O. Box 2104,
Chongqing, People's Republic of China

В МИРЕ НАУКИ
2R/copy 24R/year $70/(abroad)
Editorial correspondence:
MIR Publishers
2, Pervy Rizhsky Pereulok
129820 Moscow U.S.S.R.
Subscription correspondence:
Victor Kamkin, Inc.
12224 Parklawn Drive,
Rockville, MD 20852, USA

TUDOMÁNY
98Ft/copy 1,176Ft/year 2,100Ft/(abroad)
Editorial correspondence:
TUDOMÁNY
H-1536 Budapest, Pf 338
Hungary
Subscription correspondence:
"KULTURA"
H-3891 Budapest, Pf. 149
Hungary

العلوم
1KD/copy 10KD/year $40/(abroad)
Editorial, subscription, advertising correspondence:
MAJALLAT AL-OLOOM
P.O. BOX 20856 Safat,
13069 – Kuwait

Advertising correspondence all editions:
SCIENTIFIC AMERICAN, Inc.
415 Madison Avenue
New York, NY 10017
Telephone: (212) 754-0550 Telex: 236115



WATER-OXIDIZING CLOCK is a cyclic mechanism that supplies electrons to the P680 chlorophylls in the photosystem II reaction center. As each photon is absorbed by the P680, the clock advances by one S state, or oxidation state, and thereby releases one electron (e^-). When the clock reaches S_4, it spontaneously releases an oxygen (O_2) molecule and reverts to the S_0 state to close the cycle.

plications for the mechanism of the water-oxidizing clock, although their interpretation depends on whether the released protons come from the water molecules directly or from some other source, such as the polypeptides that bind the manganese atoms. If the protons come from the water, then the water molecules must be undergoing some chemical changes prior to the S_4 state. Conversely, if the sequentially released protons come directly from the polypeptides (and are later replaced by protons from the water molecules), then no water oxidation occurs until the final S_4-to-S_0 transition. The protons' immediate origin has not yet been determined.

Regardless of the source of the protons, it now seems likely that the higher S states (particularly S_2) accumulate some net positive charge. It is possible that a negatively charged ion may be needed to stabilize this positive charge, which could explain the observation that ions such as chloride are essential to keeping the water-oxidizing clock running. Seikichi Izawa of Wayne State University in Detroit was one of the first to demonstrate that chloride ions can turn on the water-oxidizing clock.

In collaboration with Herbert S. Gutowsky and his colleagues at the University of Illinois at Urbana-Champaign in 1982, we began to apply NMR techniques to monitor the binding of chloride ions to photosynthetic membranes. In early studies there, Ion C. Baianu, Christa Critchley and one of us (Govindjee) showed that chloride ions associate with and dissociate from isolated chloroplast membranes freely and rapidly. These findings led us to speculate in 1983 that the binding of a negatively charged chloride ion might be linked with the arrival of a positive charge on the water-oxidizing clock from $P680^+$ and that the release of the chloride ion might coincide with the release of protons.

NMR experiments by Christopher Preston and R. J. Pace of the Australian National University in Canberra suggested that chloride ions bind more tightly in the S_2 and S_3 states than in the S_0 and S_1 states. This finding is consistent with the more positively charged character of the higher S states. X-ray spectroscopy data collected by Klein and his colleagues indicate that chloride does not bind directly to the manganese atoms in the lower S states.

Peter H. Homann of Florida State University and his associates have suggested that chloride probably binds to positively charged amino acids on the proteins of the clock. Working with Gutowsky, we have made observations of chloride binding in photosystem II complexes from spinach. Our measurements indicate that several chloride ions bind to the clock and that they seem to be divided between two major binding sites: one near the manganese, perhaps on the D1 and D2 polypeptides, and the other on the 33-kilodalton polypeptide.

All these experiments suggest that the function of chloride ions in the water-oxidizing clock may be to expedite the release of protons from water. In doing this, the chloride ions may increase the efficiency of the water-oxidation reactions, or they may stabilize the charged manganese ions in the higher S states, or they may do both. The role of chloride is still controversial; it may turn out that chloride organizes the photosystem II proteins into a stable structure.

Another ion, calcium (Ca^{2+}), is essential for both the oxidation of water and the operation of the photosystem II reaction center, and it also appears to be intimately involved with the function of chloride. Experiments in several laboratories suggest that calcium ions can functionally replace two of the polypeptides at the bottom of photosystem II that are involved in the production of molecular oxygen. It has

production of molecular oxygen. It has also been observed that the removal of calcium ions seems to block both the turnover of the water-oxidizing clock (by interrupting the S_3-to-S_4-to-S_0 transition) and the fast reduction of $P680^+$ to P680.

It seems likely, therefore, that calcium has a structural or regulatory role in photosystem II. Calcium has been shown to play an important part in controlling a wide variety of proteins in other biological systems: it switches the activity of the proteins on and off and maintains their three-dimensional structure. The calcium ions in photosystem II may put the polypeptides of the water-oxidizing clock into the correct functional conformation.

The elaborate mechanism that makes oxygen during photosynthesis is only one small part of the full photosynthetic pathway in oxygen-producing organisms. Although the general details are similar among all photosynthetic species, significant differences have arisen in the course of evolution.

Most analyses indicate that the differences between the photosystem II of cyanobacteria and that of plants are relatively minor, which suggests that cyanobacteria are ancestors of, or otherwise closely related to, plants. The differences between the reaction centers of cyanobacteria and those of many other photosynthetic bacteria are much more pronounced, revealing a clear division in the evolutionary pathway. More detailed studies of the photosystems by molecular genetics, X-ray crystallography and spectroscopy will undoubtedly refine understanding of the evolution of life.

FURTHER READING

RELEVANCE OF THE PHOTOSYNTHETIC REACTION CENTER FROM PURPLE BACTERIA TO THE STRUCTURE OF PHOTOSYSTEM II. Harmut Michel and Johann Deisenhofer in *Biochemistry*, Vol. 27, No. 1, pages 1–7; January 12, 1988.

MECHANISM OF PHOTOSYNTHETIC WATER OXIDATION. Gary W. Brudvig, Warren F. Beck and Julio C. de Paula in *Annual Review of Biophysics and Biophysical Chemistry*, Vol. 18. Annual Reviews, Inc., 1989.

PHOTOSYSTEM II: FROM A FEMTOSECOND TO A MILLISECOND. Govindjee and Michael R. Wasielewski in *Photosynthesis*. Edited by Winslow R. Briggs. Alan R. Liss, Inc., 1989.

PHOTOSYSTEM II, THE WATER-SPLITTING ENZYME. A. W. Rutherford in *Trends in Biochemical Sciences*, Vol. 14, pages 227–232; June, 1989.

SCIENTIFIC AMERICAN

is now available to the blind and physically handicapped on cassette tapes.

All inquiries should be made directly to RECORDED PERIODICALS, Division of Associated Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

Want to brush up on a foreign language?

With Audio-Forum's intermediate and advanced materials, it's easy to maintain and sharpen your foreign language skills.

Besides intermediate and advanced audio-cassette courses—most developed for the U.S. State Dept.—we offer foreign-language mystery dramas, dialogs recorded in Paris, games, music, and many other helpful materials. And if you want to learn a *new* language, we have beginning courses for adults and for children.

We offer introductory and advanced materials in most of the world's languages: French, German, Spanish, Italian, Japanese, Mandarin, Greek, Russian, Portuguese, Korean, Norwegian, Swedish, and many others.

Call or write for FREE 32-p. catalog.

AUDIO-FORUM
Room M218, 96 Broad Street
Guilford, CT 06437
(203) 453-9794

SCIENTIFIC AMERICAN CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, Utah 84104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $1.50 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/89. $1.50 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the authors' risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 3187, Harlan, IA. 51593. Telephone inquiries: 1-800-333-1199, U.S. only; other 515-247-7631/32. The date of the last issue on subscriptions appears on each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, a mailing label of a recent issue) as well as the new one.

Name

New Address	Old Address
Street	Street
City	City
State and ZIP	State and ZIP

SCIENCE AND BUSINESS

Progress by Degrees

New superconductors inch toward applications

Conductors in flux, antibodies abound, dialing for dollars, the stock market

The race to commercialize ceramic high-temperature superconductors has settled into the steady pace of a cross-country marathon. Although mammoth projects such as high-temperature superconducting trains are not likely soon, investigators are making steady progress toward more modest goals. In spite of some publicized difficulties in turning discovery into practice, "the party is not over," emphasizes Sungho Jin of AT&T Bell Laboratories.

Researchers in the U.S. and Japan can now routinely make high-quality superconducting thin films and bulk materials, a task that seemed arduous less than a year ago. They are also learning to control magnetic "flux creep," the problem that generated such ominous press reports last year.

Magnetic flux creep can occur when investigators try to pass large currents through bulk superconductors at temperatures above a few kelvins in the presence of a strong magnetic field. Flux vortices in the material are pushed by the current and so inhibit the flow of current through the superconductor. Now, however, several research groups, including those at AT&T Bell, have shown they can retard flux creep by adding minor defects to the superconducting crystal. These inclusions "pin" the flux and let current pass through the material.

"Our part is to find a practical way of introducing the defects," Jin says. He and his colleagues have created crystal defects by decomposing yttrium1-barium2-copper4 (1-2-4) into yttrium1-barium2-copper3 (1-2-3). Another AT&T Bell team, led by R. Bruce van Dover, used fast neutron and proton beams to create useful deformities in crystals of 1-2-3 compounds. Jin and van Dover have achieved current densities of 100,000 and 600,000 amperes per square centimeter, respectively.

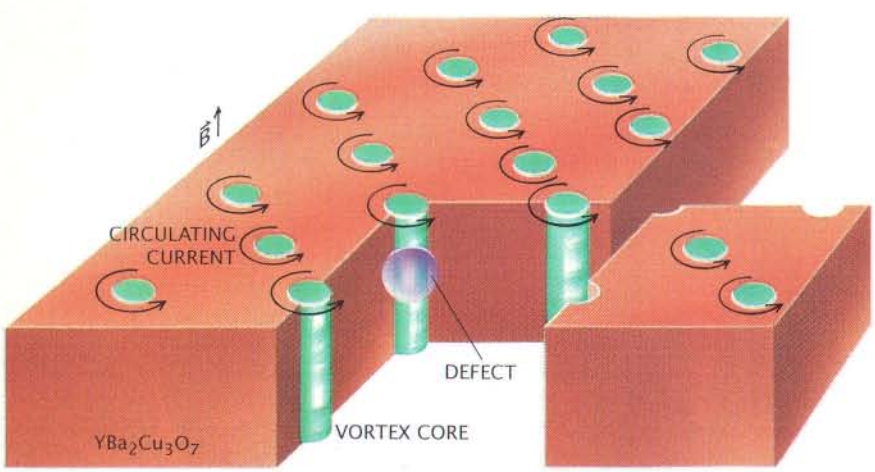
Researchers concede that processing those materials into wires, tapes and the like is still challenging. Many investigators are skeptical of reports of success. For instance, late last year, Kenichi Sato of Sumitomo Electric announced that he had developed a ceramic superconducting wire based on a bismuth-lead-strontium compound that could be spun into wire tens of meters long and that had a current-carrying capacity of 25,000 amperes per square centimeter. He has predicted that usable wires capable of carrying up to 100,000 amperes will be available within one year.

"Even if Sato has 10 meters [of superconducting wire], the problem is the quality along the length," says Hiroyasu Ogiwara of Toshiba. Commercial users, moreover, will need several kilometers of wire, observes Keiichi Ogawa of Japan's National Research Institute for Metals. "What we really want is to have a tough, long wire that could produce a big magnetic field," he says. Because such a magnetic field would put enormous stress on the superconducting wire, the material must have good mechanical properties, Ogawa notes, adding that "perhaps it will take a long time for the Sumitomo guy to solve that problem."

Still, a few simple applications incorporating high-temperature superconductors are making tentative debuts. Japan's Institute of Physical and Chemical Research (RIKEN) and Mitsui Mining and Smelting have jointly built small, cylindrical yttrium-barium-copper ceramic shields that block ambient magnetic fields, according to Hiroshi Ohta of RIKEN. Such shields can be used to help improve measurements made by conventional low-temperature superconducting quantum interference devices (SQUID's) of the faint magnetic fields produced by the brain or the heart.

So far, the largest Mitsui shield measures 15 centimeters wide and 35 centimeters long—large enough to study a rat's heart. Ohta says Mitsui plans to build shields large enough to encompass an entire person once it finds enough funding to build a massive furnace for annealing the superconductor. As soon as Mitsui signs up about 10 customers and secures some government assistance, Ohta is confident the project will roll forward.

In the U.S., AT&T Bell and the British-based ICI Advanced Materials have constructed prototypes of radio-frequency and microwave cavity resonators made of yttrium superconductors. Such resonators, which generate about 100 times less noise than con-



MAGNETIC FLUX CREEP can be diminished by introducing defects into a superconducting crystal. An external magnetic field, B, induces vortices (areas of magnetic flux surrounded by circulating current) in the crystal, which align themselves in a hexagonal matrix. A current passed through the sample pushes the vortices—the flux-creep effect—dissipating energy and so inhibiting the current. Defects in the crystal "pin" the vortices. Source: R. Bruce van Dover, AT&T Bell Laboratories.

ventional devices, could improve the sensitivity of radios and radars and increase the number of channels on cellular phones.

Farther behind are SQUID's made from high-temperature films. IBM researchers are achieving raw sensitivities in experimental devices comparable to those of commercial, low-temperature devices. Exploiting this sensitivity to make useful measurements, however, remains a challenge. Although IBM has no plans to sell SQUID's, the San Diego-based Biomagnetic Technologies is developing a high-temperature SQUID it hopes eventually to commercialize.

Researchers working to build active digital devices—such as microprocessors—must traverse a rocky terrain. Such devices would consist of an insulator sandwiched between layers of superconducting films. Unfortunately, high-temperature superconductors react with many other materials; those that are compatible often cannot tolerate the hot, oxygen-rich environment needed to form the superconducting layers. Etching circuit patterns into the films is also tricky and can easily degrade the superconductor's current-carrying properties. "Silicon is easy because there is only one component," whereas superconductors have five or six elements, Ogawa points out. "The question is, Can we fabricate the basic building-block devices?" asks Randy W. Simon of TRW.

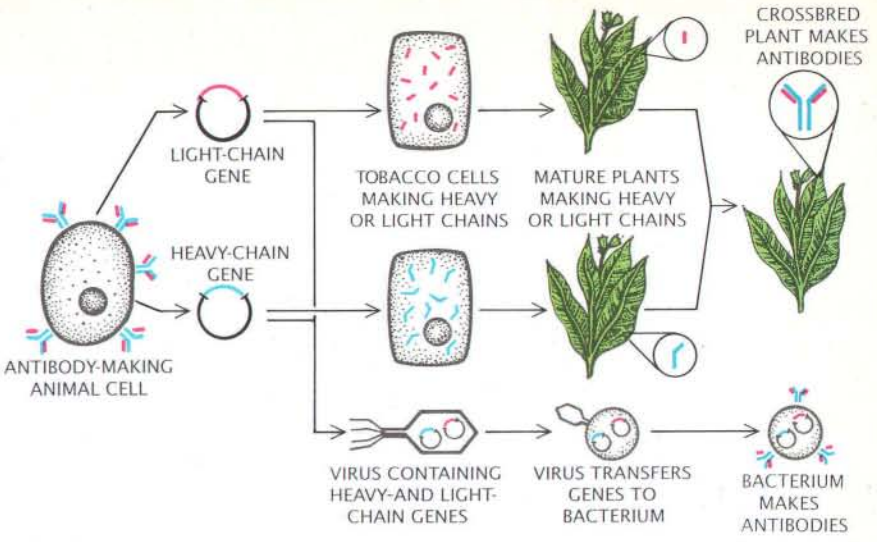
Although there are hurdles, workers are pleased with what they have already achieved. "The rate of progress has been unprecedented," Simon observes. Investigators of high-temperature materials have filed some 1,300 patent applications in the U.S. and 4,800 in Japan.

Nevertheless, a government-sponsored team of U.S. investigators who recently toured superconductivity laboratories in Japan and produced a "JTEC" report observed that a notable characteristic of their thin-film projects was the "strong commitment to materials synthesis projects even in the absence of short-term device or applications goals." U.S. research groups, in contrast, are pushing hard to develop applications. Moreover, competition within Japan is so intense that investigators there prefer to collaborate with U.S. counterparts rather than with each other.

High-temperature superconductivity, Ogawa observes, gives Japanese researchers an ideal opportunity to prove that their theoretical and basic research is worthy of Nobel-prize status. Adds Koichi Kitazawa of the University of Tokyo: "From now on, Japan will be at almost the same level as other countries in basic science."

Japanese researchers remain undeterred by the long investment needed to develop applications. "High-temperature materials are only two years old," Kitazawa says, "and everybody is expecting 'the baby' to run a 10-second 100-meter race. We need 10 years or so to really know its future."

—Frederick S. Myers and E. Corcoran



MAKING MONOCLONAL ANTIBODIES could be made easier by two new techniques developed at the Research Institute of Scripps Clinic. Antibodies consist of heavy chains* (blue) *and light chains* (red), *which are encoded by separate genes. In the "plantibody" method the genes for the heavy and light chains are isolated and transferred separately into tobacco plant cells. These cells and the mature plants that they become produce heavy or light chains. If a heavy chain–making plant is crossed with a plant making light chains, then one quarter of the resulting plants make both chains, which are assembled into complete antibodies. A second technique packages millions of pairs of heavy- and light-chain genes from antibody-making cells in viruses. These viruses infect bacteria, which produce antibodies.

Antibody Bonanza

Can two new techniques shake up biotechnology?

In La Jolla, Calif., the surf offshore cannot compare with the waves the Research Institute of Scripps Clinic has been making in the biotechnology industry lately. Last November Andrew Hiatt and his colleagues announced in *Nature* that they had produced antibodies in tobacco plants—a neat trick, because only animals make these disease-fighting proteins naturally. A month later Richard A. Lerner, William D. Huse and their associates presented a bacterial technique in *Science* for producing and screening millions of varieties of monoclonal antibodies far more quickly than is now possible.

The practical significance of these discoveries may be far-reaching. Monoclonal antibodies bind selectively to specific molecular targets; they are in demand as a component in pharmaceutical products such as drug-delivery systems. The new techniques could slash the price of monoclonal antibodies dramatically.

Moving from the laboratory to the market, however, may prove harder than making newspaper headlines. The scientific community's response to Hiatt's successful introduction of antibody genes into tobacco plants has been enthusiastic; the business response, more tempered. Johnson & Johnson, which supports biomedical research at Scripps and markets several products based on monoclonal antibodies, has already declined to license Hiatt's discovery. A decision by PPG Industries is pending.

Johnson & Johnson sources say the work did not fit into the company's current research strategy. Some makers of monoclonal antibodies suggest that "plantibodies" are still too far from commercial development and question whether the technique would solve production problems.

Undeterred, Hiatt points to three potential applications for plantibodies. In pharmaceuticals, plantibodies might be less likely to provoke an immune reaction in human beings

than might conventional antibodies, which are made in mouse cells. Crops containing plantibodies against plant pathogens might be more disease-resistant. Finally, plants with plantibodies against pollutants might also be able to extract toxic chemicals from groundwater.

Hiatt also stresses cost advantages. Today monoclonal antibodies are made in cultures of special cells called hybridomas, which consist of mouse immune-system cells fused with tumor cells. Because the antibody yields are small, the cost is high—typically about $5,000 per gram, according to some industry sources. Extrapolating from the yield of antibody protein he has gleaned, Hiatt estimates that a gram of plantibodies might cost only about 10 cents.

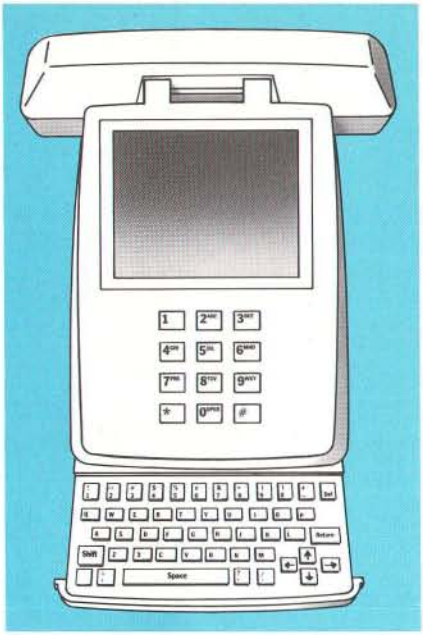
He is quick to acknowledge, however, that this calculation is highly simplistic: it ignores the costs of purifying the plantibodies from the plants, for example. Others argue that while it might cost thousands of dollars to make a few grams of a novel antibody, the unit costs plummet when manufacturers produce larger quantities. "You don't have to go to plants to make monoclonal antibodies cheaply," insists Harvey J. Berger, president of research and development at Centocor, an antibody manufacturer in Malvern, Pa.

Berger's observation seems particularly true in light of Lerner's work with bacteria. Making antibodies with bacteria is not new, but Lerner's technique is. After isolating and copying large numbers of different genes for antibody components from immune-system cells, he and his co-workers paired the genes at random. These new combinations were packaged in viruses, which infected bacteria and introduced the antibody genes. The genetically engineered bacteria then synthesized antibodies.

Whereas with hybridomas it may take years to produce and screen a few thousand different monoclonal antibodies to find a molecule with a desired specificity, Lerner can make and screen millions of new antibodies from bacteria in less than a week. He believes that his technique will lend itself to producing monoclonal antibodies derived from the human immune system, which could be safer or more effective in drug preparations. Although Centocor already makes human antibodies, Lerner maintains that his approach will be more efficient.

Much of the work on Lerner's approach was started at Stratagene, a five-year-old La Jolla-based company. Joseph A. Sorge, chief executive officer of Stratagene and co-inventor of the new technique, reports that the company's subsidiary, Stratacyte, will sell laboratory kits to researchers and will license the technology. Several companies and venture capitalists have already contacted him, Sorge says. Scripps will also offer opportunities to license the technique.

Hiatt continues his research on plantibodies. If he can demonstrate their competitive advantages, plantibodies may yet bear fruit. —*John Rennie*



ENHANCED TELEPHONE developed by Citibank can display consumers' bank balances and other data on screen. It even makes routine phone calls.

Banking Futures

Somewhere in Gotham City a bank is testing technology

Deep in the heart of midtown Manhattan, in the basement of an ordinary office building, Citibank is testing the consumer-banking technology of the future.

For more than a dozen years Citibank, the nation's largest bank, has quietly used its "Lab" to try out ideas for new consumer services. The lab "gives us the freedom to fly," says Lawrence D. Weiss, who directs Citibank's development division. In the 1970's the lab helped to give birth to the bank's automated teller machines, called Customer Activated Terminals (CAT's). More recently it has helped design Citibank's newest home-banking device, an enhanced telephone, playfully referred to as "ET."

The lab, which for many years was a secret even within the bank, is one way Citibank has shown its zeal for technology. (Citibank allowed SCIENTIFIC AMERICAN to tour the lab on the condition that we would not reveal the location.) Consultants at McKinsey estimate that the bank spent about $1.5 billion in 1988 on technology systems from computers to telecommunications—about 13 percent of the banking industry's outlay.

Although the nitty-gritty engineering of new devices is typically done by Citicorp's subsidiary, Transaction Technology, Inc., Weiss relies on the lab to find out whether customers will like the innovations. "This is theater," declares Gloria K. Mendez, who directs the lab. The lab secures a prototype, redecorates its rooms to simulate the appropriate setting—say, a bank office or a living room—and invites volunteers to try out the device.

Citibank wastes little time making prototypes; when the bank began experimenting with early CAT designs, lab vice president Donald G. Fannon, Jr., hid behind the display and handed out play money. The one-way mirrors and array of video cameras surrounding the lab's rooms are real, though. Citibank workers watch and record consumer reactions to a device, then modify the technology accordingly.

The ET's that Citibank plans to roll out in early 1990 gestated for more than two years in the lab. Weiss says the bank was exploring ways to encourage financial transactions at home. Banks have found consumers reluctant to do their banking via personal computers. (Citibank's Direct Access has about 40,000 subscribers, far fewer than the bank first hoped but more than its competitors have.)

Weiss's development team turned to a telephone-based system. The lab tested scores of designs before concluding that the device should look and work like a conventional phone first and a banking tool second, Fannon recalls. The messages that flash on the screen were also tested carefully. People were most comfortable when the messages were perceived as friendly—but not giddy.

"We've made it very familiar," Weiss says. A user can make a conventional phone call or by pressing the "#" key can summon a menu of options on the embedded screen. She can then speed-dial a call, add numbers to an internal directory by punching in the details on a slide-out keyboard or get in touch

with her bank account. Selection of the bank-account option brings up another menu with choices that include finding out an account balance, paying bills, seeing current Citibank interest rates and so on. Once a transaction is finished, the hallmark Citibank phrase appears on the screen: "Thank you. It's always a pleasure to serve you."

The bank says it is not using technology to supplant its work force, pointing out that in the past decade it has hired more people even as its spending on technology has increased dramatically; nevertheless, that possibility causes some jitters. Weiss emphasizes, however, that technology—such as the new phone—helps Citibank employees shift from simply responding to customers' needs toward actively anticipating them.

Meanwhile, down in the lab, Mendez and Fannon are busily redecorating the rooms. —*E.C.*

THE ANALYTICAL ECONOMIST

Taking stock

As the bull market continues to limp along, complaining about the stock market has become a national pastime. Some corporate executives argue that they are shackled to their quarterly returns; even if the company is sound, low returns may spark a flurry of selling by shareholders. Others grumble that the market undervalues their companies and leaves them vulnerable to takeovers. Investors fret that the market has become so volatile that they are unlikely to make a profit. What purpose does the market serve these days? Is it doing an efficient job?

Long before the New York Stock Exchange was officially established in 1817, merchants traded company stocks (and government bonds) as a means of raising and making money. Investors swapped their money for equity, or ownership, in a company. Companies still look to stock markets for capital, although not as often as they once did. (According to Securities Data Company, companies raised in excess of $20.7 billion by issuing common stock last year as compared with $29.9 billion in 1988.)

Economists define a broad role for the stock market: it is a mechanism for allocating a scarce resource, namely, capital. Because investors continually look for a good return on their money, they will move funds out of companies that are faltering and into ones that are well managed. This flexibility sets a price for every company and makes it easier or more difficult for the firm to raise additional funds.

If the market is "efficient," then such prices truly reflect the value of the company. An inefficient market, on the other hand, wastes money by pushing up the stock prices of companies that have lackluster futures. It also cheapens the value of better firms. Unfortunately, there is no consensus among economists about how to determine whether the market is indeed efficient.

According to the benchmark definition written by Eugene F. Fama of the University of Chicago, an efficient market "correctly uses all available information" to determine the appropriate stock prices. A negligible transaction cost for trading stocks and free access to information are two necessary conditions for an efficient market.

Several changes during the past 15 years have made the market increasingly adept at moving capital around. In 1975 the U.S. abolished a rule that required brokers to charge fixed fees for their services. By freeing brokers to negotiate (and thus lower) their fees, the change has made it cheaper to trade large blocks of stock. Computers have also speeded up trading, bringing more information to investors and enabling them to act quickly on the news. Whereas in the past an investor might have learned that a company lost a bid for a new contract but then decided it was too costly to bother selling that stock, today he might sell his stock immediately, says Gregg A. Jarrell of the University of Rochester.

As a result, proponents of the efficient-market theory say, it may be appropriate for a company's stock to rise or fall dramatically when the market learns something new. When investors learned in late October that after 30 years Texas Instruments had finally won a Japanese patent for the integrated circuit, they pushed up TI's stock by almost 30 percent. "If you have good information, the market sets a fairly precise number on the stock price," Jarrell maintains.

Even the bidding wars that take place during a corporate takeover and send a company's stock skyrocketing can be explained, Jarrell says. He argues that the stock price is elevated by the bidder's plans for reorganizing the company. If such bids fall through, the stock quickly sinks almost to its original level, he says, because none of the proposed restructuring takes place.

Other economists remain unconvinced that the market responds in such a rational way. Robert J. Shiller of Yale University argues that stock prices are often bid up or down by investors who simply follow a market fad and get carried away.

As evidence of the effects of fads on the market, Shiller points to what he sees as excessive volatility of prices in the market. In theory, investors assess the value of a stock based on a calculation of its "present value." This standard calculation is essentially a summation of the expected dividends from a stock, discounted by an interest rate. (The discount rate acknowledges that inflation and uncertainty make the future dividends less valuable than payments today.) If an investor decides that the present value of a company exceeds the current selling price, he should buy. Moreover, since dividends are somewhat predictable, Shiller argues, changes in trading patterns should be smooth. Instead he sees erratic fluctuations in the data and so believes that the market is not behaving rationally.

Shortly after the Dow Jones Industrial Average plunged 190 points on October 13 of last year, Shiller surveyed about 100 market professionals on what they felt caused them to sell. Some 77 percent thought that the changes in the market that day were more the result of psychology and emotion than of news about a change in stock-market fundamentals.

But the economists do agree on one point: the market is not significantly more volatile now than it has been in the past. A recent study by G. William Schwert of the University of Rochester shows that the percentage changes in the Dow during the 1980's have been unremarkable—except, of course, on October 19, 1987, when the Dow fell more than 20 percent. As a percentage, the more recent October 13 fall did not even rank among the 25 largest drops in the stock market's history, he says. Schwert suggests that press reports about the absolute fall in the Dow (190 points on October 13) have fanned concerns, whereas reporting the drop as a percentage (more than 6 percent) would not have.

The issue of market efficiency is far from resolved. If the stock market is efficient, however, executives who complain that their companies are undervalued should take a long, close look at the way they run their business. —*Elizabeth Corcoran*

Progress in Gallium Arsenide Semiconductors

The compound is not a candidate to supplant silicon. Nevertheless, its speed and optical capabilities have spawned fast-growing applications in computing and communications

by Marc H. Brodsky

Today's global age of electronics is built on a miniature foundation of microscopic circuits engraved on silicon chips. The current success and continuing promise of silicon in consumer, commercial, industrial and military electronic systems has prompted those who work with the material to offer a tongue-in-cheek criticism of another promising semiconductor, gallium arsenide. "Gallium arsenide," they say, "is the technology of the future, always has been, always will be."

After almost 30 years as the technology of the future, gallium arsenide has begun to make a place for itself, not by supplanting silicon but by complementing it in new applications. The inherent advantages of the material lie in the speed with which electrons move through it, in weak-signal operations and in the generation and detection of light. These advantages suit it for roles in computing, television reception and the optoelectronic transmission of data through optical-fiber networks (a technology also known as photonics). Gallium arsenide light-emitting diodes and lasers used in visual-display technologies and audio-disk players already account for more than $1 billion in sales annually. Hundreds of thousands of satellite-receiving dishes that use gallium arsenide detectors are sold every year, and high-speed circuits using gallium arsenide transistors are projected to reach a similar turnover in a few years. In an economy and society that depend on the rapid exchange of information as well as on the processing of it, many silicon-dominated processors will require a considerable admixture of gallium arsenide components in order to do their jobs.

MARC H. BRODSKY, a physicist, has conducted research on the fundamental properties and applications of semiconductors at the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y. He had headed IBM's Advanced Gallium Arsenide Technology Laboratory until his recent appointment as director of technical planning in the company's research division. Brodsky received his B.A., M.S. and Ph.D. degrees from the University of Pennsylvania. He has long had an interest in science education and has served on professional and community organizations for the improvement of educational opportunities.

Gallium arsenide technology has largely followed the course of development charted earlier by workers in silicon. Since the invention of the transistor in 1948 by John Bardeen, Walter H. Brattain and William B. Shockley of Bell Telephone Laboratories, researchers have tried to improve semiconductors in two ways. First, physicists and electrical engineers seek materials that can switch on and off more quickly and perhaps perform other tasks, such as the detection and generation of light. Indeed, it was toward these ends that gallium arsenide—which does not occur in nature—was formulated in the 1950's by Heinrich Welker of Siemens Laboratories. He also investigated closely related semiconductors derived from elements in the columns of the periodic table adjacent to silicon and germanium, the constituents of the earliest transistors.

Second, engineers refine the techniques by which semiconductors are manufactured. This work requires that the semiconductors' chemical and physical traits be specified and that compatible auxiliary materials and processes be developed for the fabrication of insulators, conductors, external connectors and other essential components. Semiconductors have to be carefully purified, combined with other substances in precise ratios and formed into perfect crystals; flaws introduced during the fabrication of transistors and circuits must be smoothed away without compromising desired electronic qualities. None of these tasks is easy even now; they were harder still in the early years, when a new materials science had to be created from fundamental studies in physics, chemistry, metallurgy and other disciplines. I shall try to guide the reader through these intertwining areas of physics, engineering, materials and electronics to show why gallium arsenide is both promising and challenging to bring to market.

Gallium arsenide's most promising property is the great ease with which electrons move through it: when all else is equal, gallium arsenide circuits are faster at equal or lower power than are silicon circuits. Because gallium arsenide consumes less power, it produces less waste heat that must be drawn from the circuit. This quality is particularly valuable because there is a trade-off between a semiconductor's speed and power.

An engineer must look at the question of speed in the context of a device, not a pure crystal of an element or compound. Today several kinds of transistors serve as the essential switching elements in electronic circuitry [*see illustrations on pages 58 and 59*]. Calculations are carried out or data processed by effecting changes in these devices. Such changes can proceed no faster than the switching speed, the time it takes an electron to traverse the semiconducting region under the control of electrical signals from another part of the circuitry.

A semiconductor's switching speed

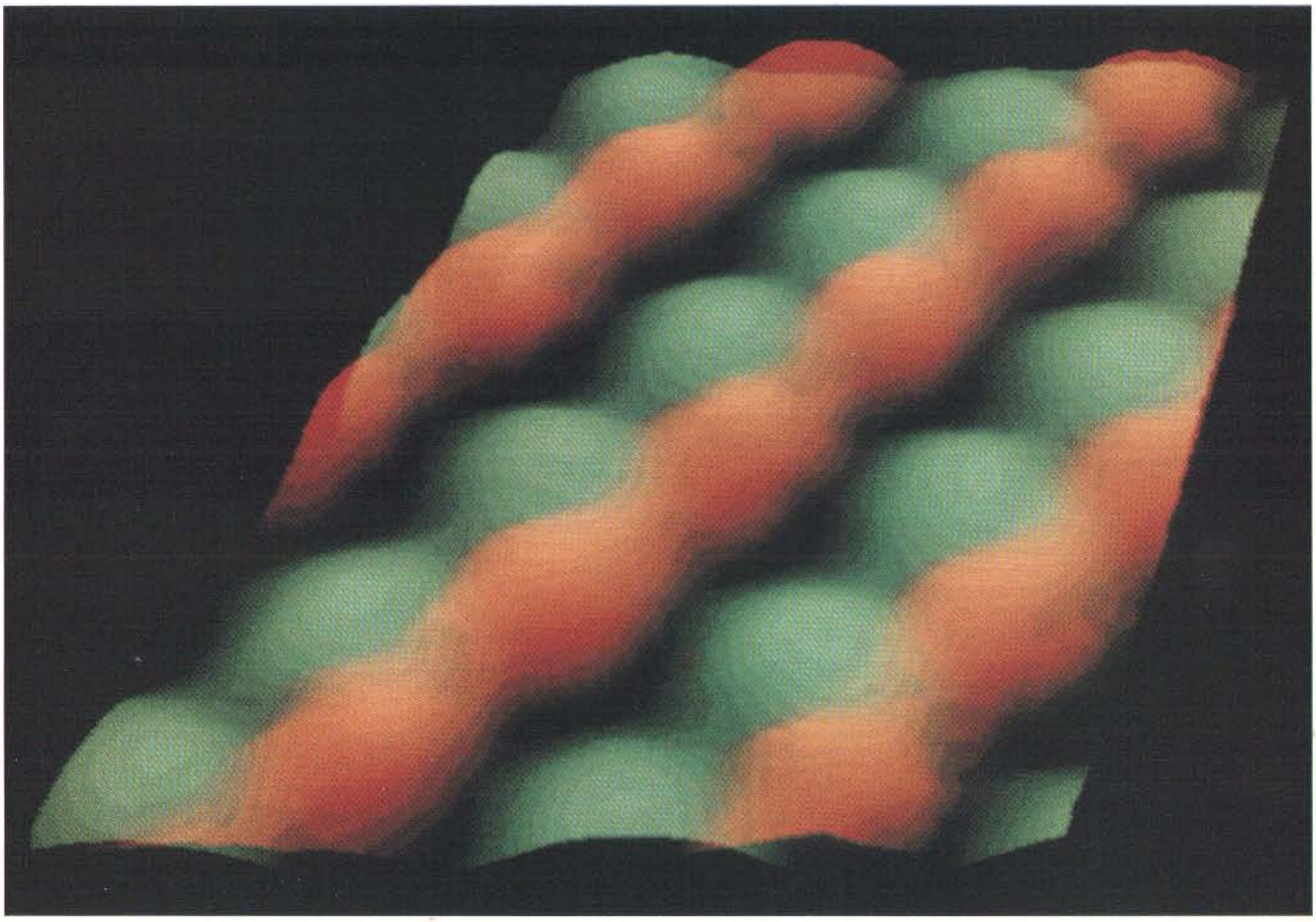
depends on the average velocity it allows an electron to attain—about one million or more centimeters per second—as it encounters numerous obstructions while traveling through a transistor. After many collisions with atoms, ions and each other, the electrons acquire a characteristic distribution of velocities that is determined by the electric field driving them and by the way that the residual impurities and the semiconductor's constituent atoms scatter them. Electrons ricochet in all directions, often losing energy in the process, slowing their net flow in the direction of the electric field.

A mechanical analogy helps to explain how physical properties can influence the movement of an electron in a semiconductor. Two different semiconducting materials can be represented by tubes that are lined with stationary and vibrating obstacles and tilted equally with respect to the ground. The obstacles are the scattering mechanism, the tilt provides a gravitational field that corresponds to an electric field and the balls dropped into the tubes represent electrons. Here, the switching speed corresponds to the time it takes the balls to reach the bottom of the tubes. Electron mobility, on the other hand, is the ease with which the balls evade the obstacles. In part, we can think of this nimbleness as if it reflected the size of the balls: a smaller, quicker ball hits fewer obstacles.

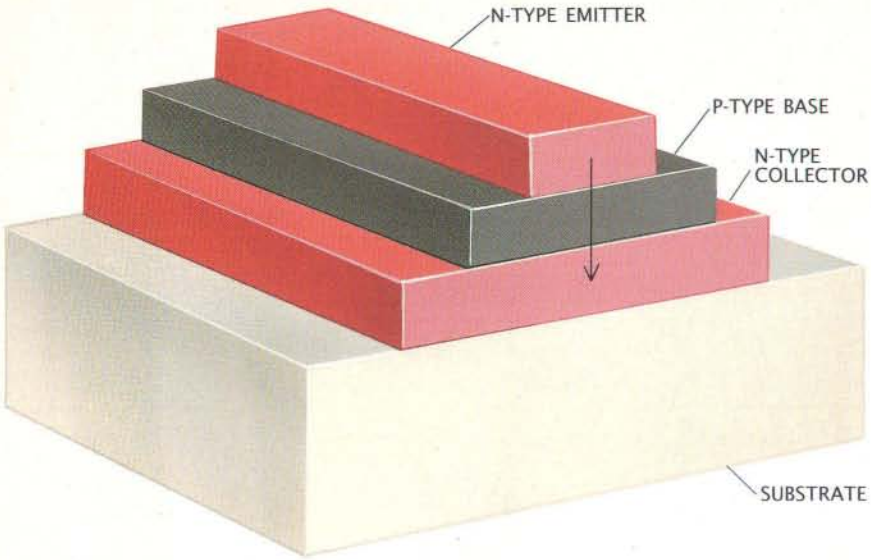
In a semiconductor the electrons move through an array of constituent atoms arranged in a crystalline lattice. Because the conduction electrons are shared by all the atoms, the lattice has the electronic character of a single tube for the passage of electrons. Arrays formed by gallium and arsenic atoms, shown in the scanning tunneling electron micrograph below, attract moving electrons less strongly than do arrays of silicon atoms. Physicists therefore regard electrons as having a smaller effective mass in gallium arsenide than they do in silicon. Since other factors are also contributing to higher mobility, it follows that electrons in gallium arsenide can generally reach higher velocities over a given distance and travel farther between collisions than can electrons in silicon.

One might stop scattering altogether by shortening the critical pathways to less than the average distance between electron collisions. That would ensure that most electrons hurtle through the critical switching area on ballistic paths [see "Ballistic Electrons in Semiconductors," by Mordehai Heiblum and Lester F. Eastman; SCIENTIFIC AMERICAN, February, 1987]. But practical application of this principle is not expected until years after the more conventional gallium arsenide transistors have found their place on the technological menu.

The ball-and-tube analogy of electron collisions pertains only to acceleration under electric fields weaker than 10,000 volts per centimeter. Stronger fields elicit additional physical interactions that produce so-called saturation as the electrons become effectively heavier; this narrows or even reverses gallium arsenide's edge over silicon. In fact, gallium arsenide tends to operate at lower optimal voltages than does silicon, making the



NESTED gallium (*green*) and arsenic atoms (*orange*) appear in this scanning tunneling micrograph made by Randall M. Feenstra and Joseph A. Stroscio of the IBM T. J. Watson Research Center. Images correspond to electronic contours of each atom.



BIPOLAR TRANSISTOR switches a circuit connected to its emitter and its collector, two regions containing trace amounts of electrons supplied by donors (*n*-type doping). These areas are separated by a controlled region, called the base, which contains holes (*p*-type). When current is applied to the base, electrons injected by the emitter pass through to the collector, switching the circuit on. These devices can be fabricated directly on wafers of silicon by photolithographic and other growth and processing methods. Gallium arsenide is made into bipolar devices by depositing it in three layers; one or more may be alloyed with varying proportions of aluminum.

best circuits of the two kinds of semiconductors hard to connect to one another. Voltage compromises must be made when the two coexist.

Mobility is also important for high-frequency, low-noise operation. Noise, the random fluctuations in voltage that determine the weakest usable signal, can be minimized at high frequencies by maximizing electron mobilities, both in the transistor and in its connections to the rest of the circuit. The lower-noise operation of gallium arsenide circuits turns out to be particularly valuable for the detection of television and microwave signals.

The other major superiority of gallium arsenide over silicon lies in the much greater ease with which the separations between its electronic bands, or energy levels, can be engineered. Such "band-gap engineering" can produce versatile optoelectronic capabilities and more flexible transistor designs. An electronic band, which defines the range of energies an electron can have, is the broadened form of an energy state of an electron orbit in an individual atom. In a pure semiconductor the valence band (which contains the electrons providing chemical bonding) is essentially filled; the next higher level, the conduction band, is essentially empty. Mobile charges in these bands are created by a process called doping, which is the precise addition of trace impurities to the host semiconductor. Regions that have electrons in the conduction band, called *n*-type semiconductors, are made by doping with atoms that function as electron donors; *p*-type regions are formed by creating positively charged holes (missing electrons in the valence band).

The energy difference between the top of the valence band and the bottom of the conduction band is called the band gap. It is larger in gallium arsenide than in silicon, but it can be narrowed or widened by judicious substitution (alloying) with other atoms. If aluminum, for example, is fully substituted for gallium to create aluminum arsenide, a much wider band gap is attained. Partial substitutions produce band gaps directly proportional to the fraction of aluminum in the alloy. Other valuable alloys are created by substituting some indium for gallium, some phosphorus for arsenic, or both at the same time.

Alloys of aluminum gallium arsenide have spacings between atoms that match those of pure gallium arsenide so closely that the two materials can be fitted together atom by atom without defects. Very thin layers of two or more alloys can be alternated to create heterojunctions, structures whose band gaps vary from layer to layer. One example is the semiconductor superlattice invented by Leo Esaki and his colleagues at IBM. In this structure, alternating layers of aluminum gallium arsenide and gallium arsenide are deposited onto a gallium arsenide substrate. Electrons moving parallel to the layer planes in such a multidecker sandwich would normally be confined to the lower band-gap layers of gallium arsenide. For them to move at right angles to the laminations, or layers, they must penetrate or go over the aluminum gallium arsenide band-gap barriers. By varying the number, width and composition of the layers, one can manipulate the electronic and physical attributes of the semiconducting heterojunctions.

Perfect crystal-to-crystal growth is essential in the exploitation of gallium arsenide's flexible band gap. Unfortunately, many gallium arsenide alloys that have desirable electronic properties exhibit unsuitable atomic spacings; their crystalline lattices do not mesh with each other or with gallium arsenide. If two mismatched crystals are laminated, rows of atoms are occasionally skipped at the joining faces, creating faults that can propagate through a layer and destroy its electronic usefulness. This effect limits the range of substances that can be conjoined. Some lattice mismatch can be accommodated by keeping one of the crystalline layers very thin, but this requirement also restricts design.

Big changes in band-gap size can sometimes be accommodated by employing a different substrate. For example, indium phosphide substrates are used in the indium-gallium arsenide-phosphide lasers that are optimal for long-distance optical-fiber communications.

Sometimes one can avoid direct mismatches between two crystalline layers that have desirable electronic qualities by separating them with specially tailored superlattice buffers. The buffers are made from various alloys whose crystals have atomic spacings that are intermediate between those of the active layers; they gradually absorb the mechanical strain over several layers. The buffers also can insulate circuits from any residual defects in the gallium arsenide substrate. Greatly improved lasers and transistors have been built using superlattice buffer layers.

In 1979 Raymond Dingle and his colleagues at Bell Laboratories demonstrated a design for heterojunction field-effect transistors that placed a gallium arsenide channel under a layer of aluminum gallium arsenide. They realized that after a dopant atom in

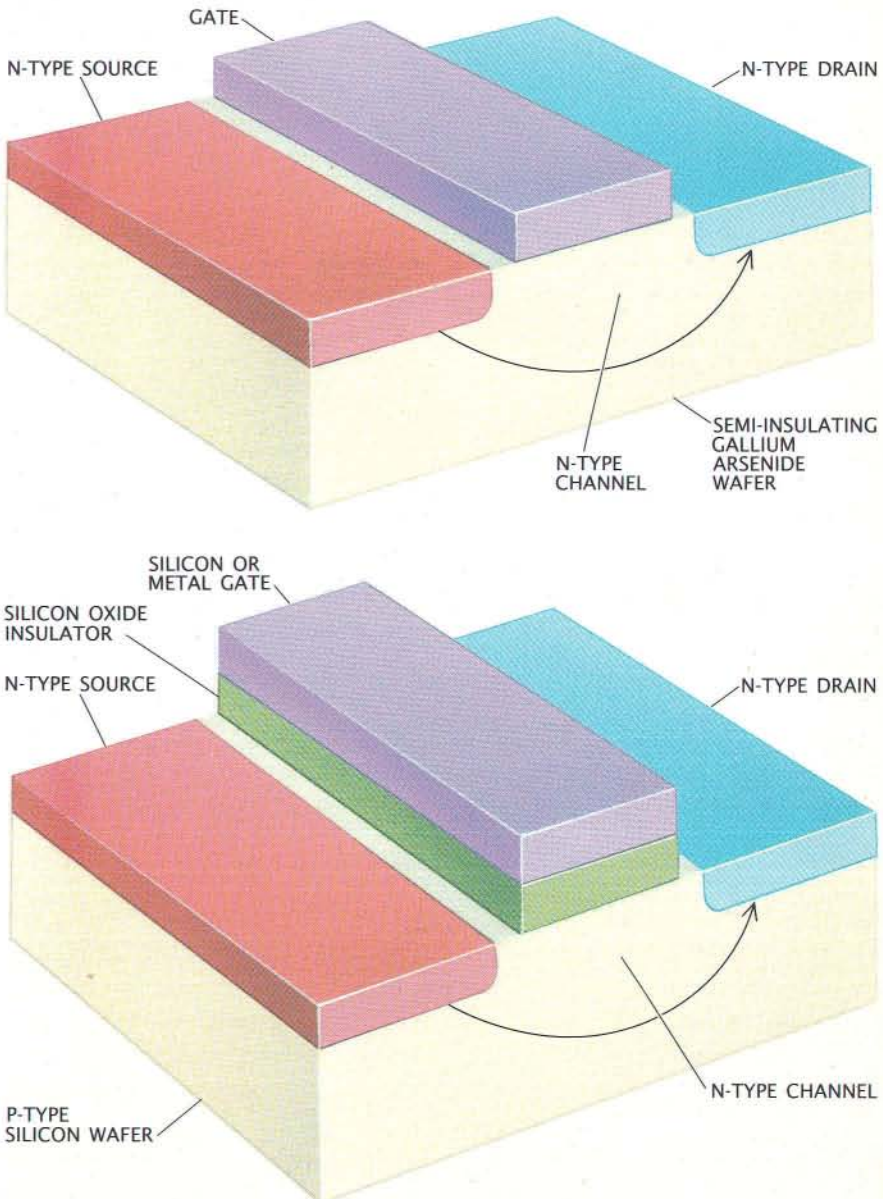
a gallium arsenide layer has donated an electron to the conduction band it becomes a positive ion that will scatter other electrons. So the workers physically separated the donor atoms from the gallium arsenide by placing dopants in an adjacent layer of aluminum gallium arsenide. Electrons donated by the dopants move to the lower conduction band of the nearby gallium arsenide layer. They therefore move faster than would have been the case had the ionized donor atoms remained in the channel, thus blocking the electron's paths. This technique, called modulation doping, was quickly incorporated into gallium arsenide field-effect transistors at Fujitsu Laboratories in Japan and Thomson-CSF in France. It increases electron mobility by only about 20 percent at room temperature. When cooled to 77 kelvin (77 degrees Celsius above absolute zero), transistors based on modulation doping offer about three times the mobility of conventionally doped devices.

Another material aspect of gallium arsenide that may help speed circuits is the capacity of its wafers to serve as semi-insulators. Wafers are the substrates on which devices and circuitry are formed. Substrates of silicon, because of the material's smaller band gap and greater amount of activated residual (unwanted) dopants, are always to some extent electrically conductive, and they therefore add capacitance, or electrical drag, that reduces the speed with which electrons traverse a circuit. Gallium arsenide's wider band gap allows it to be prepared in semi-insulating form either by keeping the substrates utterly free of active dopants or by incorporating special self-compensating dopants that almost completely cancel the effect of residual dopants. Such advantages over silicon may not persist at levels of integration beyond several thousand circuits per chip, at which point the capacitance between the numerous closely spaced wires connecting the circuits becomes the crucial limit to signal speed.

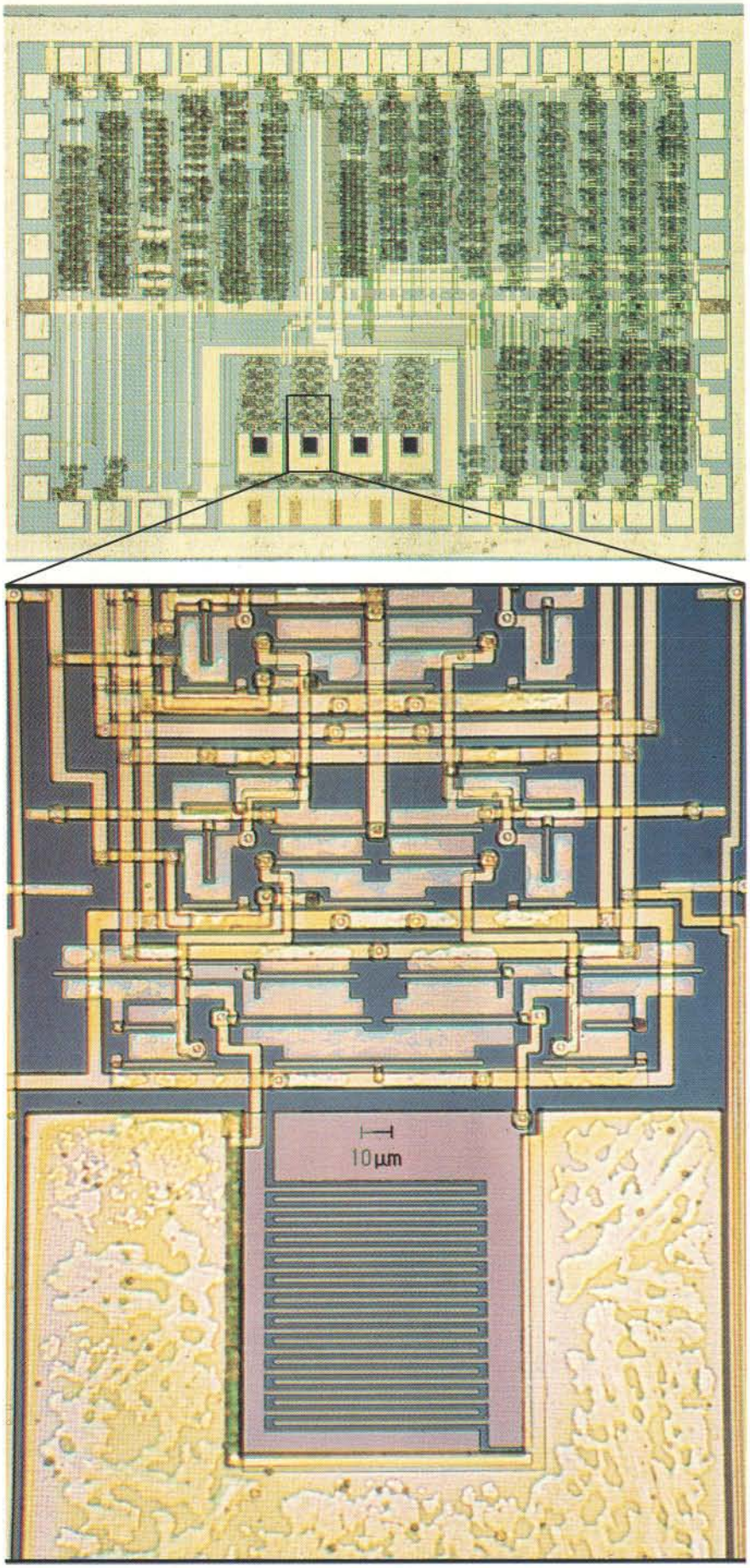
In addition to electron mobility and band-gap flexibility, gallium arsenide's third and most dramatic advantage over silicon is its capacity to radiate and detect near-infrared radiation. In gallium arsenide the potential energy of an electron moving from the conduction to the valence band can easily be given up as a quantum of electromagnetic radiation, or photon. The same reaction in silicon generally requires a nonradiative reaction, such as a collision, in order to conserve momentum. This difference in band-gap properties explains why gallium arsenide can support optoelectronic functions and silicon cannot.

A gallium arsenide radiation source consists of a diode in which two regions are oppositely doped, giving the *p*-type one an abundance of holes (in the valence band) and the *n*-type one an abundance of electrons (in the conduction band). Voltage applied to such a *p*/*n* diode injects holes and electrons into the boundary between the two regions, creating excess populations of electrons and holes. When an electron and a hole combine, annihilating one another, the electron's band-gap energy is released as a photon.

This wonderfully efficient electron/hole recombination (which under certain circumstances can convert most of the electrical energy into light) can be used in a straightforward manner to create a familiar electronic component: the light-emitting diode (LED). LED's of aluminum gallium arsenide or gallium arsenide phosphorus are seen



FIELD-EFFECT TRANSISTOR controls the passage of electrons through a channel under a controlled electrode, the gate. When a positive charge is applied to the gate, it attracts electrons to a thin region at the top of the substrate, creating a conducting *n*-doped channel. That channel conducts electrons from an *n*-type region called the source to an *n*-type region called the drain, closing the circuit to which they are connected. In the silicon version of this device (*bottom*), that element's high-quality native oxide is used as an insulating layer between the gate and the substrate. A common gallium arsenide design (*top*) instead applies the gate directly to the substrate.



in many electronic displays as flashing red or yellow dots, respectively. Tens of millions of them are made every year.

A laser diode generates more concentrated radiation. The faces of the crystal that forms the device are perfectly parallel and act as semitransparent mirrors. Light reflects back and forth through the recombination region, where it stimulates the emission of radiation that has the same wavelength and directionality. The resulting beam is highly coherent. The original semiconductor diode lasers were made from gallium arsenide *p/n* junctions in independent experiments at IBM and General Electric in 1962. Advanced band-gap engineered designs depend on structures containing layers of different compositions. For example, gallium arsenide lasers emit radiation at near-infrared wavelengths because the band gap of gallium arsenide corresponds to an energy just below that of visible photons. Alloying gallium arsenide with one or more elements (aluminum, indium, phosphorus) shifts the gap to higher or lower values, pushing the output farther into the infrared or moving it into the visible region.

Equally important is the capacity of gallium arsenide and its alloys to detect light by reversing the reaction underlying LED's and laser diodes. The resulting photodetectors can be tuned to a given wavelength by means of the same band-gap engineering techniques that are used to tune laser diodes. Because gallium arsenide photodetectors are so efficient, they can respond far faster than silicon ones. They have the additional virtue of being easily integrated into high-speed gallium arsenide electronic circuitry.

OPTOELECTRONIC CHIP of gallium arsenide receives infrared light on its photodetector (*detail*) at the rate of one billion bits per second and converts it into a weak electrical signal. The photodetector and electronic amplification and processing circuitry are on the same chip. A wire too short to pick up unwanted cross talk from neighboring circuits connects the detector to the first of several stages of amplifiers. Additional circuits containing thousands of transistors sort the data into eight-bit packets, called bytes, which are then fed to digital computers. This chip forms part of a three-chip unit that receives and transmits data with unprecedented packing of optoelectronic functions; it was made by IBM's gallium arsenide research unit.

Such integration is a major goal for economic and functional reasons. First, like all miniaturization, it would greatly reduce the unit costs of the device (allowing a single wafer to produce scores of chips, each containing thousands of circuits). Second, by placing a photodetector extremely close to the circuit that first amplifies it (the "front-end" circuit), one can design the connection between the two elements so as to minimize the antennalike pickup of unwanted signals from neighboring circuits, called cross talk—a major problem in circuits connected to nonintegrated detectors by conventional wire links. Beyond the integration of optical and electronic functions lies the more speculative prospect of chips in which one optical signal directly modulates another. Many such devices have been proposed, but all are far from realization.

Besides the ability to generate and detect light, other advantages make gallium arsenide the potential technology of choice in specific applications. For example, its wide range of operating temperatures and great resistance to high-energy radiation make it valuable for automotive and military applications, respectively.

It is not enough to identify gallium arsenide's properties and devise ways of exploiting them; one must also make the products themselves at high quality and low cost. This problem of manufacturing technology, to which we now turn, is in many ways the most formidable of all.

Gallium arsenide, like silicon, forms electronic elements in accordance with rules set by its physical and chemical characteristics. These rules are complicated by the need for many components in an integrated circuit that are not made of semiconductors. Among them are the metal connectors that link the elements, the insulators that separate them, the resistors and capacitors that control the flow of current and the dopants that supply electrons or holes.

One of the most serious shortcomings of gallium arsenide and its related alloys is the lack of a usable native oxide, such as that which silicon forms when heated in air. Silicon oxide constitutes an electronic and mechanical seal that has a variety of applications. In field-effect transistors [*see illustration on page 59*], silicon oxide provides the insulation between the gate and the channel. In bipolar transistor circuits, silicon oxide provides insulation between adjacent transistors. Silicon oxide can also be exploited as a tool in the fabrication of transistors and circuits on chips; it provides a protective mask through which windows can be cut to allow reactive chemicals to etch patterns, deposited metal to build conducting films and added dopants to activate particular regions.

Gallium arsenide technologists instead must form insulators and masking structures by other, often less convenient means. A common design for a field-effect transistor places a metal gate directly in contact with gallium arsenide, which produces a controllable channel by means of an effect called Schottky barrier formation [*see illustration on page 59*]. Other designs grow thin crystalline layers of aluminum gallium arsenide as insulatorlike barriers. Whereas aluminum gallium arsenide provides flexibility for modulation doping, as described above, neither this technique nor the Schottky barrier allows the range of operating voltages afforded by silicon oxide insulators.

Another major disadvantage of gallium arsenide and related compounds derives from the very fact that they are compounds. Whereas defects in elemental silicon can be annealed by heating the crystal so as to shake any wayward atoms into line, in gallium arsenide this process competes with the selective vaporization of arsenic. Defects arise from many of the processing steps in the fabrication of integrated circuits. In particular, dopants are typically added by accelerating ions to implant them into the semiconductor. In the case of silicon the resulting damage can be removed, and all of the implanted dopants can be eased into their proper places in the crystal by annealing the material at nearly 1,000 degrees C for several minutes (during which the silicon oxide coat prevents the dopants from boiling off). But ion-implanted gallium arsenide cannot be annealed so successfully, even when the temperature is kept as low as 800 degrees C. Special precautions to retain the arsenic and the dopants achieve only partial success: only 90 to 95 percent of the originally implanted dopants are generally activated.

New techniques based on annealing cycles of only a few seconds and special capping layers (which form a seal) are now being tested, but so far no practical method of carrying annealing to completion has been devised for gallium arsenide. Therefore, device characteristics tend to vary across a chip, which, along with other factors, places limits on the level of integration that can be achieved. Whereas a silicon chip of a square centimeter can now hold more than a million transistors, the state-of-the-art gallium arsenide chip can accommodate only tens of thousands of components. Because the cost of processing a gallium arsenide wafer is the same as or even more than that of processing a silicon wafer (both of which contain many scores of chips), this comparatively low level of integration becomes a distinct disadvantage. A further economic handicap is imposed by the smaller size of today's gallium arsenide wafers: usually they measure only three to four inches in diameter. Silicon wafers currently used in manufacturing have a diameter of five to eight inches.

Applications of gallium arsenide have been mainly limited to devices for which the relatively high unit cost is affordable because the function is unique. The semiconductor is most commonly used in front-end, high-speed receivers, in which fast response and low noise are needed, and in optical generation, for which there is no substitute material. Applications in digital circuitry are beginning to follow in the highest-performing computers, and there are proposals for the use of gallium arsenide in future microprocessors.

Perhaps the most familiar front-end application of gallium arsenide detectors is seen at the focus of satellite dish antennas. Communications satellites employ microwaves of up to 12 gigahertz, a spectral region in which gallium arsenide's speed clearly leaves it without peer. Not only can gallium arsenide convert these wavelengths to clear electronic signals, it can also amplify the weak initial electrical signals almost noiselessly. Silicon amplifiers can switch that rapidly only by using transistor connections that add enough noise to drown out the weak signals. The advance of integration should eventually make it economical to expand the use of these front-end devices into such commercial products as the television tuners that were recently demonstrated by several companies in Japan and Europe. These tuners can be expected to deliver clearer pictures in areas where reception is marginal.

Gallium arsenide's most important growing application is undoubtedly in the photonic transmission of information. Light propagation in fibers can carry much more information and carry it farther than can electrical signals in ordinary metal wires. The linking of

digital and other systems by optical fibers is increasing at a rapid pace. Transcontinental and transoceanic fibers are already in use, and fiber is spreading through the telephone and cable-television systems toward home and office. Optical fibers are also being used to extend the distance and speed of computer connections. Such linkages will increasingly become part of the digital-processing system itself. Some analysts believe that optoelectronics will have as great an effect on society in the coming years as digital computing has had already.

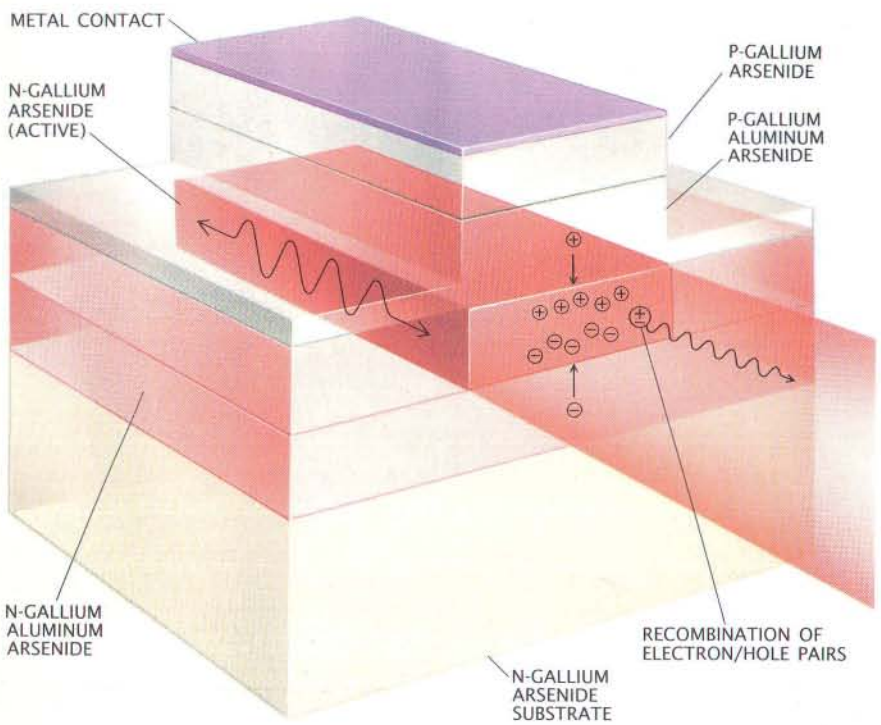
The laser diode is to the gas laser as the transistor is to the vacuum tube; the same analogy holds for LED's when compared with incandescent bulbs. In each case the semiconductor is smaller, more efficient and cheaper to make and operate than is its bulky analogue. The margin of advantage is not measured in a few percentage points but in orders of magnitude. These differences make possible such new applications as the compact-disk player. This consumer product employs an aluminum gallium arsenide laser to read data encoded as marks on a rotating disk. An aluminum gallium arsenide laser operating at higher power is an essential component of a related device, a computer data-storage unit based on an optical disk.

Such solid-state lasers are particularly apt candidates for band-gap engineering. This technique can be used to control precisely the wavelength that the laser emits. Visible light in the red wavelengths is emitted by lasers made from layers of lattice-matched aluminum-gallium-indium-phosphide grown on gallium arsenide substrates. These devices are being tested as replacements for cumbersome gas lasers in such applications as the bar-code scanners that check groceries at cashier counters or automobile parts in factory assembly lines. Indium-gallium arsenide-phosphide lasers are widely used in long-distance communications because their output can be tuned to the infrared wavelengths near 1.3 or 1.55 micrometers, which are the least easily absorbed in optical fibers.

So far workers in optoelectronics have largely concentrated on improving the operation of discrete devices rather than on integrating them into a single unit that can be embodied in a chip. In part, this focus is the heritage of the pioneering role played by long-distance telephone companies, which needed the optical links in applications where the unit cost was of secondary importance, because it could be spread over many telephone conversations. Hence, commercial devices employ separate chips for the lasers, the detectors and the transistors.

Progressive linking of computers by optical fibers will require large numbers of much less expensive optoelectronic devices. Such devices can be made at affordable unit prices only through techniques of large-scale integration. Optoelectronic links will eventually have to transmit data at one billion bits per second or more, rates that in principle can be achieved by either field-effect or bipolar gallium arsenide transistors. Advanced silicon transistors of the bipolar design can also serve this function. Yet gallium arsenide field-effect devices have become the technology of choice because they dissipate less power and can later be extended to still higher operating speeds.

In 1979 Amnon Yariv and his colleagues at the California Institute of Technology built the first interconnected transistor and laser on a gallium arsenide chip. Matsushita and NEC Central Research Laboratories in Japan have also made similar indium phosphide-based devices; exploratory work in various compounds continues in many other laboratories. The major practical challenge consists in overcoming the mismatches between the processing steps by which optimal versions of both the transistors and the lasers can be fabricated on a single chip.



LASER DIODE injects holes from a *p*-type layer of gallium aluminum arsenide downward into an active layer of gallium arsenide; an *n*-type layer of gallium aluminum arsenide injects electrons from below. The excess populations of holes and electrons annihilate one another, releasing the energy difference between the electron's conduction band and the hole's valence band as a quantum of light. Quanta moving parallel to the mirrored vertical faces are reflected into the active layer, where they stimulate the emission of quanta having the same wavelength and directionality.

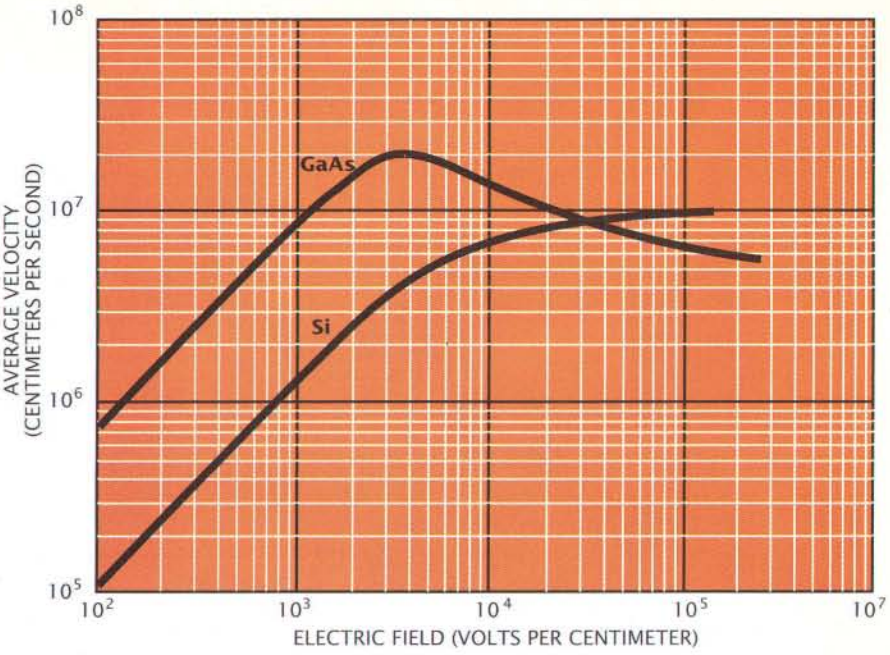
My colleagues in the advanced gallium arsenide technology laboratory and other IBM facilities, working at three centers in New York State and one in Switzerland, recently built and packaged a trio of chips that transmit one billion bits per second. None of this speed will go to waste: transmission links must be about 10 times faster than the computers they connect. That is because the links transmit data in series, whereas computers process data in eight-bit batches, or bytes (with two bits added to check for errors in transmission). The fastest computers already produce a stream of data that outstrips the carrying capacity of copper cable at distances beyond 200 meters (where the closely spaced electrical signals begin to blur). Optoelec-

tronic systems, with their vastly greater bandwidth, are replacing cable at progressively shorter distances; eventually they will be used to route information within the computer itself.

Our circuit is a package consisting of three chips and a connecting bundle of optical fibers. The first chip, a gallium arsenide transmitter, serializes bytes arriving at a rate of 100 million per second and feeds them to the second chip, an aluminum gallium arsenide laser array. One of its four lasers flashes the signal over an optical fiber at one billion bits per second to the third chip, a gallium arsenide receiver, one of whose four built-in photodetectors converts the signal to electronic pulses. These pulses directly feed into gallium arsenide transistors, which amplify them [*see detail in illustration on page 60*]. Other circuits on the receiving chip then deserialize the signal into bytes. At either end of the optoelectronic link, the gallium arsenide transmitter and receiver chips connect to silicon circuitry that actually processes the bytes as part of a computer. But as data-processing rates increase, gallium arsenide can be expected to work its way into the digital circuitry of the systems being interconnected.

That circuitry lies at the heart of digital computing, a field of paramount importance in which gallium arsenide has found little application to date. Most common digital applications require circuits having higher levels of integration—and thus lower unit circuit costs—than gallium arsenide has yet achieved. Gallium arsenide's speed advantage has been exploited only in some "high-end" digital applications in mainframe computers and supercomputers, which emphasize performance more than cost. (Supercomputers achieve clock speeds of up to 200 megahertz, about six times faster than those of the fastest deskside machines.) Moreover, these specialized applications give silicon less of an edge in miniaturization, because the most powerful silicon transistors generate so much heat that their density must be carefully limited. Circuits must therefore be scattered over a number of chips, thereby creating a new bottleneck in the form of interchip-transmission delays. Computer engineers try to minimize such delays by packaging related chips closely together.

Silicon will continue to serve as the stuff of digital computing unless gallium arsenide technology can be pushed to higher levels of integration and packaging compatibility. Interchip delays must also be minimized so that any advantages on the level of the chip can be realized in the entire system. Supercomputer projects in Japan and the U.S. are using modest levels of gallium arsenide circuit integration and sophisticated packaging techniques. Clock speeds of as high as 300 megahertz are envisioned for the Cray-3 supercomputer, the first of its kind to be based on gallium arsenide.



AVERAGE ELECTRON VELOCITIES of gallium arsenide are more than five times greater than those of silicon in fields weaker than about 1,000 volts per centimeter.

Not all the competitive disadvantages of gallium arsenide are technical. Although gallium arsenide technology has employed many methods that were pioneered in silicon, such as photolithography, this advantage is offset by the enormous past investment in silicon, which therefore tends to get the nod over gallium arsenide wherever its performance is not notably the worse. Furthermore, because most of the new investment in semiconductor research still flows to silicon, the established material continues to present a formidable moving target. Gallium arsenide's advantages in speed, low noise and optoelectronics must become valued enough to overcome the challenge of manufacturing economics unless or until the technology catches up to silicon.

A technological path by which gallium arsenide may one day invade general digital computing is suggested by reduced instruction set computing (RISC), a new technique that speeds processing on chips by using a reduced number of elements to execute specialized functions. Several companies producing workstations have projected the development of RISC to speeds that only gallium arsenide could handle. Today's RISC chips are based on silicon transistors that process about 35 million instructions per second. Existing silicon technology can accommodate a tripling of this rate. Speeds of beyond 100 million instructions per second can be achieved only by gallium arsenide transistors. Economic feasibility in workstations demands that at least 40,000 transistors be packed on a chip, a level of integration that has already been achieved by several companies in California's "Gallium Gulch"; the companies can be expected to market fully validated circuit designs within a year or two. Perhaps then, when computers, computer links, televisions and compact disks all contain gallium arsenide, we will be able to say that the technology of the future has finally arrived.

FURTHER READING

MOLECULAR BEAM EPITAXY AND HETEROSTRUCTURES. Edited by Leroy L. Chang and Klaus Ploog. Kluwer, 1985.

SEMICONDUCTOR DEVICES, PHYSICS AND TECHNOLOGY. S. M. Sze. John Wiley & Sons, Inc., 1985.

IEEE GAAS IC [GALLIUM ARSENIDE INTEGRATED CIRCUITS] SYMPOSIUM. IEEE, November 6–9, 1988.

Food Sharing in Vampire Bats

Two nights without a blood meal and a vampire bat starves to death—unless it can solicit food from a roostmate. A buddy system ensures that food distribution among the bats is equitable

by Gerald S. Wilkinson

At night—long after most visual predators have stopped prowling—vampire bats emerge from their roosts and take to the wing, flying low across the landscape in search of warm-blooded prey. Within an hour or two, having found appropriate victims and fed on their blood, the bats return to the roost to sleep, feed their young and interact with nestmates.

Until recently little was known about either the behavior or the life history of the common vampire bat, *Desmodus rotundus*. For many years biologists were more interested in the animal's physiology than in its social organization, which was thought to be relatively simple. A number of recent studies, however, reveal that vampire bats are remarkably social: females cluster together during the day but at night reassort themselves, creating a fluid social organization that is maintained for many years. Moreover, it is now known that long-term associations among females enable bats to regurgitate blood to one another on a regular basis and so significantly increase their chances of survival.

The reason for regurgitation behavior was revealed in studies carried out more than 15 years ago by Brian K. McNab of the University of Florida, who showed that a vampire bat will die if it fails to feed for two nights in a row. After 60 hours without food it loses as much as 25 percent of its weight and can no longer maintain a critical body temperature. To fuel the body's metabolic engine and avoid death, individuals must consume 50—sometimes even 100—percent of their body weight in blood every night.

Yet feeding is not always easy, especially for young bats, who must learn to bite quickly and to do so without inflicting pain on their victims. I have seen horses toss their heads, swish their tails and rub against obstacles to rid themselves of hungry bats. Although the bats counter such defensive strategies by returning to the same animal (a known target) several nights in succession or by feeding sequentially from fresh wounds, from 7 to 30 percent of the bats in a cluster fail to obtain a blood meal on any given night. By soliciting food from a roostmate, a bat can fend off starvation—at least for one more night—and so have another chance to find a meal.

In 1978 Uwe Schmidt, a zoologist at the University of Bonn, presented the first evidence that females regurgitate blood to their pups. At that time Schmidt had kept bats for more than 10 years in a turret at Poppelsdorfer Schloss, an old castle that is now the main research building of the university's Zoological Institute, and had spent much of his career observing their behavior. Schmidt discovered, for example, that shortly after birth, the pups are given regurgitated blood—in addition to milk—by their mothers; he also found that on some occasions a pup will take blood from an adult other than its mother. In one case he even observed an orphaned pup who was being suckled by an adoptive parent. Food sharing of this sort, in which individuals provision other members of a group, is extremely rare in mammals; in addition to bats, only a few species—such as wild dogs, hyenas, chimpanzees and human beings—are known to display such behavior.

Food sharing appears to be altruistic: a donor bat gives up food—which might otherwise be used to ensure either its own survival or the survival of its offspring—to a recipient bat, whose chances of survival are thereby increased at no apparent cost to itself. Yet true altruism has never been documented in nonhuman animals, presumably because such a one-way system is not evolutionarily stable. The reason is that donors, who lose resources, are eventually outcompeted by recipients, who have more resources and so survive longer, produce more offspring and pass more of their genes on to the next generation. Careful studies of altruistic behavior by a number of investigators reveal that many acts of apparent altruism actually take place either between relatives (and so are a form of kin selection) or between individuals who exchange resources on a more or less equal basis, in which case they can be considered to be acts of reciprocal altruism, or reciprocity.

Having heard of Schmidt's work, I wanted to study vampire bats in their natural habitat to see whether blood regurgitation is an act of kin selection or reciprocity (or both). I set off for Costa Rica, and there, with the help of my assistants Robin Weiss, Michael L. Jones and Terri Lamp, I studied a population of *Desmodus rotundus* for 26 months between 1978 and 1983.

I hoped by observing their regurgitation behavior to see if the bats were feeding only their relatives, and

GERALD S. WILKINSON has spent more than 10 years studying the social behavior of bats, having first become interested in them while taking a field course in Costa Rica in 1978. Since then he has studied bats in South America, Africa, Australia, Southeast Asia and the U.S. Wilkinson received his B.S. from the University of California, Davis, in 1977 and his Ph.D. in biology from the University of California, San Diego, in 1984. He has been assistant professor of zoology at the University of Maryland at College Park since 1987.

COMMON VAMPIRE BAT, *Desmodus rotundus*, is found from Mexico south to Argentina and Chile, particularly in areas where the land has been converted to pasture. Females, such as those shown here, gather together in caves and hollow trees during the day, emerging only at night to search for warm-blooded prey.

therefore engaged in kin selection, or if they were reciprocally exchanging food (with either related or nonrelated individuals), and thus engaged in reciprocity. In order to prove reciprocity I needed to demonstrate that five criteria were being met: that females associate for long periods, so that each one has a large but unpredictable number of opportunities to engage in blood sharing; that the likelihood of an individual regurgitating to a roostmate can be predicted on the basis of their past association; that the roles of donor and recipient frequently reverse; that the short-term benefits to the recipient are greater than the costs to the donor; and that donors are able to recognize and expel cheaters from the system.

Vampire bats, which are common in tropical America wherever land has been converted to pasture and livestock are present, are ideal subjects for a study of this nature. For my research site I selected a cattle ranch in northwestern Costa Rica called Hacienda La Pacifica (which has since been turned into an ecological research station and is now called Centro Ecologica La Pacifica).

There I found that vampire bats, in the absence of caves, spend the day in hollow trees, where temperatures are constant, the humidity is high and it is dark even during the day. Most of the trees at La Pacifica had a single opening at their base: by lying inside the opening and peering upward with the aid of binoculars and a diffuse light source, we could observe the bats for several hours at a time. By gradually increasing the amount of light (over a period of several months), we could habituate the bats to our presence and observe their interactions with one another. The single entrance at the base of the tree offered another advantage: we could stretch a fine-mesh net in front of it and catch the bats when they emerged at night to hunt. In this way we were able to tag them and subsequently quantify individual patterns of behavior.

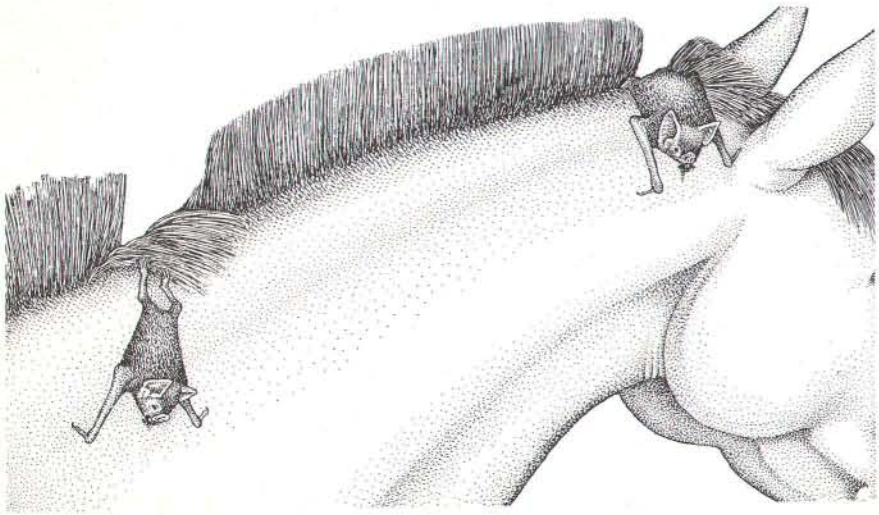
We found that the bats emerge in search of prey every night at a time that varies with the phase of the moon: if it is too bright outside, the bats wait until the moon goes down. Unlike the other two species of blood-drinking bats (the white-winged vampire bat, *Diaemus youngi,* and the hairy-legged vampire bat, *Diphylla ecaudata*), which feed mostly on the blood of birds, *Desmodus rotundus* feeds primarily on mammalian blood. The bat seems to prefer horses to cows, and it locates them by a combination of smell, sound and echolocation.

Having identified a victim, the bat usually lands on the animal's tail or mane and hangs from it while searching for an appropriate spot to bite. Specialized heat-sensitive cells in the nose help the bat find a place where the victim's blood vessels are near the surface. When such a spot is found, the bat quickly excises a small patch of skin with its razor-sharp upper incisors and begins feeding. An anticoagulant in the bat's saliva keeps the blood flowing during the 20 to 30 minutes needed to consume a meal. Then the bat, its stomach visibly swollen, returns to the roost, where it remains

FACIAL FEATURES of the vampire bat reflect its blood-based diet. The enlarged ears help the animal search for prey and navigate by echolocation: the bat emits high-pitched sounds that reflect off objects in its path, and the echoes are picked up by the ears. Smell and heat receptors in the broad, fleshy nose enable the bat to home in on a suitable victim. After finding an appropriate spot to bite, the bat excises a small patch of flesh with its razor-sharp incisors. With the aid of an anticoagulant in its saliva, the bat then laps up the blood that flows from the victim's wound.

until the next night. As the bats returned to their roosting trees just before dawn, we netted them to determine which individuals had successfully obtained a blood meal.

Our first goal on arriving at La Pacifica was to tag all the bats in our study area with lightweight bands of different colors; females were banded around the left wrist and males around the right. Each band also had a piece of reflective tape affixed to it, which enabled us to spot roosting individuals even when light levels in the tree were very low. Once the banding process was completed, we could pick a bat at random and record its behavior every 10 seconds, 100 times in succession. By working in pairs, we tagged 600 bats and accumulated more than 400 hours of such behavioral observations. Once a week we did a census of the animals in each tree to document patterns of association among bats occupying the same daytime roost. We also fitted a total of 37 bats with radio transmitters and so were able to determine the extent and degree of overlap of their foraging ranges.



COMMON VAMPIRE BATS feed on a variety of mammalian prey but seem to prefer horses and mules to other species. They usually land on the animal's tail or mane, which gives them something to hang onto during the 20 to 30 minutes it takes to drink their fill. The animal is not always cooperative; it may shake its head and swish its tail in an attempt to dislodge the bats. A bat who fails to feed for two nights in a row will die from starvation unless it can solicit regurgitated blood from a roostmate.

We found that the social organization of vampire bats is dominated by groups of from eight to 12 adult females and an equal number of pups (one for each female). Pups are born throughout the year at about 10-month intervals; females stay with their mothers, whereas males leave between the ages of 12 and 18 months, when they become reproductively mature. In contrast to some other tropical bat species, in which males defend "harems," vampire bat males defend territories. They form dominance hierarchies within hollow trees, fighting among themselves for the alpha position near the top of the tree's hollow cavity (where females frequently cluster) and defend their territories vigorously—sometimes to the death—against intruders. The intruders are males who normally roost alone or in small groups during the day in trees that are rarely visited by females.

As we tracked the bats and monitored their associations, we were surprised to find that their social organization is stable and yet fluid. The bat population in our study area could be subdivided into three groups of about 12 adult females, each of which often subdivided into smaller clusters. Although the three groups were isolated from one another, the composition of the individual clusters that made up each group varied continually. Each group had exclusive rights to a range of about six trees, and once or twice a week the females would shift roosts (sometimes carrying their pups with them to another tree), reassorting themselves in the process.

Because female pups stay with their mothers past reproductive maturity, several generations are typically found clustered together in one tree. Yet my biochemical analyses of blood samples suggest that only about 50 percent of the offspring in a cluster share the same father. Presumably

this is because females show no loyalty to a particular tree (whereas males do) and so are periodically exposed to new males, with whom they sometimes mate. In addition, for reasons that are not well understood, females switch groups from time to time (possibly because prey have become difficult to locate); on the average a new female joins a group every two years. As a result, each group consists of several matrilines, within which relatedness is high but between which it is low.

Analysis of roosting associations reveals that adult females exhibit a preference for certain other females, which cannot be explained simply on the basis of some physical feature of the roosting site. Moreover, it seems that their preference is not always for relatives but may be for nonrelatives, a finding consistent with reciprocity theory. Having determined that the bats have an affinity for one another, we needed to answer the following question: Do the bats remain together for long periods and thus have the opportunity to develop and maintain a mutual support system?

The answer appears to be yes. Rexford D. Lord, now at the Indiana University of Pennsylvania, determined the maximum life expectancy of vampire bats by counting the annual growth rings of their teeth and found that females can live for as long as 18 years. And from banding studies undertaken at La Pacifica by Theodore H. Fleming of the University of Miami in the 1970's, we knew that at least two of the females in our study area had roosted together for more than 12 years. In view of their longevity and the fact that each individual fails to feed periodically, we concluded that the bats meet the first criterion of reciprocity: not only do individuals have the ability to form long, stable relations with one another, but the opportunity to engage in food sharing is ever present among them.

Our next step was to determine whether blood is regurgitated randomly within a group or whether the females regurgitate only to close relatives or to prior roostmates, as predicted by kin selection and reciprocity theory, respectively. To do so, we needed to estimate the frequency of blood sharing under natural conditions.

During the course of our five-year study we witnessed a total of 110 instances of blood sharing by regurgitation. Seventy percent of the regurgitations took place between a mother and her pup and can therefore be thought of as parental care. The remaining 30 percent, however, involved adult females feeding young other than their own, adult females feeding other adult females and, on two occasions, adult males feeding offspring.

To determine whether or not bats regurgitate selectively, we compared the degree of relatedness between a recipient and a donor as well as their roost-association index (the proportion of times two individuals were seen together in the same cluster) to see if—on either account—the regurgitation values were higher than if the recipient were randomly soliciting from all potential donors in the roost. We found as a result that both relatedness and prior association are important predictors of an individual's response to a solicitation. Our results show that vampire bats do not share blood randomly but share preferentially with individuals who are frequent roostmates and often, but not always, related, a finding that supports both reciprocity and kin-selection theories.

The next step in our study was to test reciprocity experimentally. If reciprocal altruism does occur among vampire bats, then one might predict that individuals should aid only those in imminent danger of starvation and should preferentially repay those bats who had previously fed them. To test these predictions, we captured four adult females from our main study area at La Pacifica and four from a secondary study area at Parque National de Santa Rosa some 50 kilometers farther north. We knew from their tags that two of the La Pacifica bats were grandmother and granddaughter (related to one another by one fourth); the others were unrelated but had a high degree of roost association.

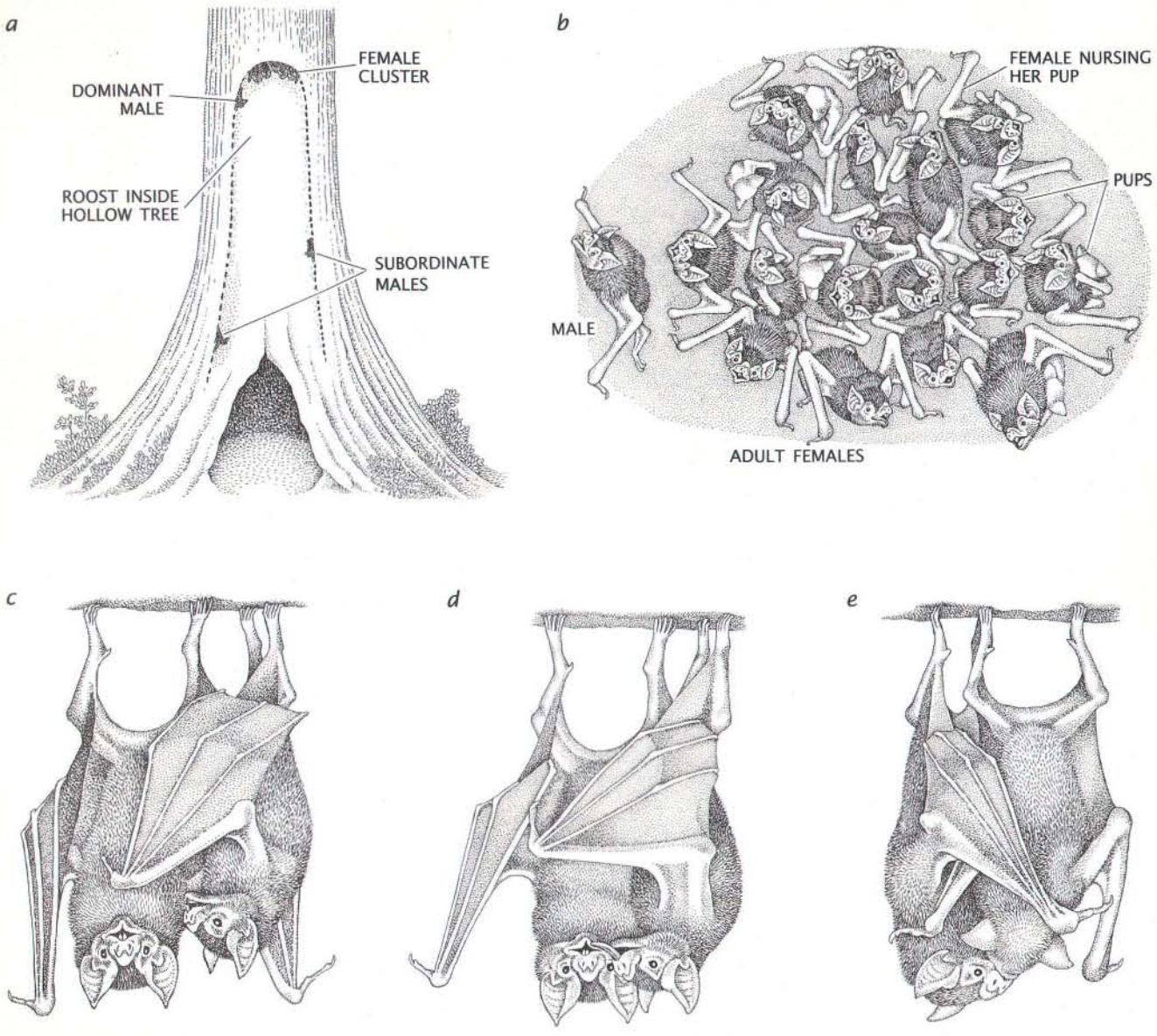
We initiated our experiment by habituating the bats to captivity and also to being fed nightly from plastic measuring bottles, which enabled us to record the amount of blood each bat would ingest at mealtime. Once the bats were at ease in their cages, we selected one each night at random and put it in a separate cage, where it was deprived of food. The next morning we returned the experimental bat to its original cage and observed its interactions with its cagemates. Our results indicate that blood shar-

STUDYING VAMPIRE BATS in their natural habitat requires that many hours be spent in an uncomfortable position (*left*). To observe the bats, which roost in hollow trees (as well as in caves and other dark places), the author and his assistants spent from two to six hours a day on their backs, peering upward with a diffuse light. Colored bands placed around the wrists of the bats (*right*) made it possible to identify individuals and follow their interactions over a five-year period.

ing nearly always took place between bats from the same population. Only once did it occur between strangers. Moreover, we found that blood sharing was not random, even among bats who had a high degree of prior association. Instead it appeared that the unrelated bats developed a buddy system, so that two individuals would regurgitate almost exclusively to each other—a strong indication that their roles reverse on a regular basis.

Another criterion of reciprocity theory is that the cost of donating blood must be small relative to the benefit it provides to the recipient. In other words, by regurgitating blood to a roostmate, the donor should be able to save its roostmate's life without substantially risking its own. To test whether or not this was true for vampire bats, we needed to measure the costs and benefits of blood sharing in two ways: directly by determining the amount of blood and the frequency of its ingestion required to prevent starvation and indirectly by estimating the effect of blood sharing on long-term survival by means of computer simulations.

From McNab's work as well as our own, we knew that a bat must consume from 20 to 30 milliliters of blood every 60 hours to prevent starvation. In addition, we knew that a bat on the brink of starvation can gain up to 12 hours of life and another chance to find food if it is given blood by a roostmate. A cooperative roostmate, who has recently fed and therefore has an elevated metabolic rate, loses less than 12 hours by donating a blood meal and so has 36 hours and two nights of hunting left before reaching the point of starvation. According to



MALE AND FEMALE vampire bats often roost in the same tree (*a*). Females cluster near the top of the cavity, some 12 or more feet from the ground, where they are guarded by a single dominant male. Two or three subordinate males occupy the same tree but roost closer to the ground. As many as 12 females, each with a young pup (the pups differ in size because births occur throughout the year), may gather in one tree (*b*). Although the composition of the roosting groups varies from day to day, some females associate for many years and regurgitate blood to one another, a behavior that is a form of reciprocal altruism. A hungry bat solicits regurgitated blood from a roostmate first by grooming (*c*), which consists of licking the potential donor under her wing, and then by licking the donor's lips (*d*). If the donor is receptive, she responds by regurgitating blood (*e*). Only bats who are close relatives or who have had a long-term association give blood to each other.

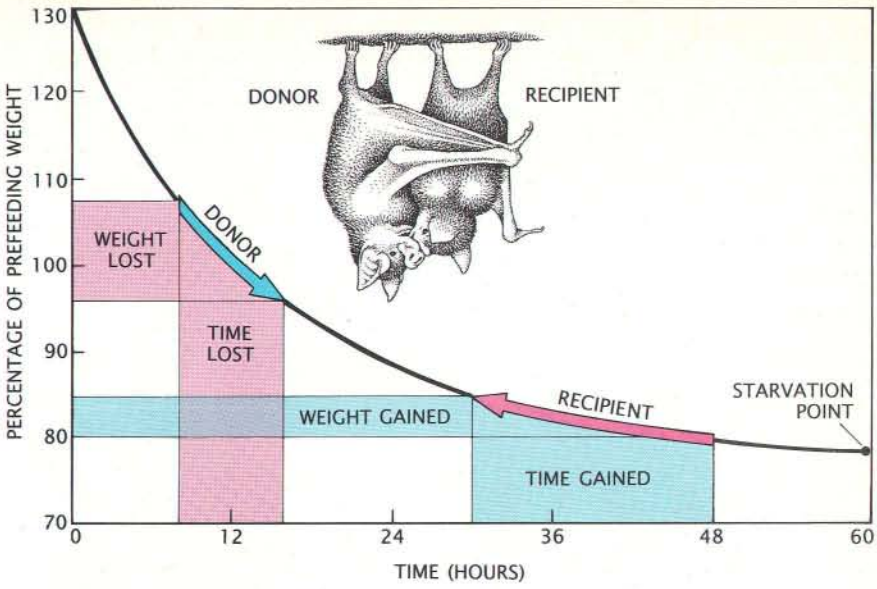
reciprocity theory, then, only bats with less than 24 hours of life remaining should be fed by their roostmates. Our experiment with captive bats, in which we withheld food for 24-hour periods, showed exactly that.

For a reciprocal-exchange system to work, it is necessary for bats to recognize one another and be able to detect and exclude cheaters. Although we have yet to prove that cheaters either exist or are excluded from the system, we have reason to believe that the bats are capable of individual recognition. To begin with, we know from our observations of captive bats that only individuals who have had a prior association will regurgitate blood to one another. It therefore seems likely that they must recognize each other in some way.

Circumstantial evidence strongly suggests that social grooming plays a role in roostmate recognition. The bats spend more than 5 percent of each day grooming and licking one another, and the behavior seems to be an important prelude to regurgitation: hungry bats frequently groom potential donors (females who have recently fed). As with blood regurgitation, grooming occurs more frequently among individuals who either are close relatives or have previously associated with one another than it does among bats who are strangers.

Additional evidence that bats recognize each other has been provided by Schmidt. By analyzing sonograms he and his students have found that the bats emit individually distinct vocalizations. Such "contact calls" often accompany grooming sessions and have the acoustic characteristics (variable frequency and low intensity) that are necessary to encode individual identity. Auditory signals (and possibly olfactory cues) of this type presumably enable individuals to recognize their long-term roostmates as well as cheaters who solicit blood but do not themselves respond to solicitation.

It seems, then, that both reciprocity and kin selection promote blood sharing among roostmates and that regurgitation is beneficial—at least in the short term—to a recipient. How might such energetically costly behavior affect overall survival rates within a population? To establish that reciprocity can persist in an evolutionary sense, one must be able to demonstrate, at least theoretically, that bats who share food with their roostmates have a higher annual survival rate than those who do not.



COST-BENEFIT ANALYSIS of blood sharing among vampire bats indicates that recipients benefit more than donors lose. The author weighed adult females returning to the roost after feeding and then weighed them every hour for the next 24 hours. An individual who had fed might return at 130 percent of its prefeeding weight (half the weight of a blood meal is lost through urination within the first hour after feeding), whereas a bat who failed to feed on two successive nights might return at 80 percent of its earlier weight. By regurgitating five milliliters of condensed blood to a hungry roostmate, the donor bat might drop from 110 to 95 percent of its prefeeding weight but lose only six hours of the time it has remaining until starvation. The recipient bat, however, gains 18 hours and so benefits more than the donor loses.

We knew from our netting studies that on the average about 30 percent of the immature bats (those younger than two years) fail to obtain a blood meal on a given night, whereas only 7 percent of reproductively mature males and females fail to feed. We also knew from field observations that the failure to feed appears to be random: all individuals within an age group are affected equally. With that information and the help of a computer, we determined that annual mortality for adults in the absence of food sharing (given that 7 percent of the adults fail to feed on a given night and that failure to feed two nights in a row will lead to death) should be about 82 percent. Because actual mortality among adult vampire bats is only 24 percent per year, we concluded that food sharing must be favored by natural selection.

Vampire bats have evolved a system of food exchange whereby they share blood with roostmates in need. Although the behavior puts the donor at risk, the recipient is more likely to survive another night. Moreover, our studies have shown that individuals who exchange blood with their roostmates gain an immediate advantage in terms of increasing their own survival and sometimes the survival of their relatives. Hence, both reciprocity and kin selection appear to be operative in this system.

Is food sharing a behavior unique to vampire bats? Not exactly. Females of several insectivorous bat species, including the bent-winged bat, *Miniopterus schreibersi*, the Mexican free-tailed bat, *Tadarida mexicana*, and the evening bat, *Nycticeius humeralis*, nurse young other than their own.

The bent-winged bat has not yet been studied in great detail. Gary F. McCracken and his students at the University of Tennessee at Knoxville have been studying the nurseries of free-tailed bats for the past nine years. Several million free-tails congregate in caves in the southwestern U.S. every summer to give birth synchronously to their pups. The young are kept in creches, where the density of individuals is as high as 40 pups per 16 square inches. The females roost elsewhere and visit their pups only twice a day to nurse them. As the females approach, the hungry pups swarm toward them; as many as four pups have been observed attempting to feed sequentially from one female.

To an observer, it appears as if females are feeding whichever pups reach them first. For that reason many investigators thought the females were a communal resource; after all, finding one's own pup amid

MEXICAN FREE-TAILED BATS rear their pups in communal nurseries, which may contain a million or more newborn bats. Despite the density of pups—as many as 40 within the space of 16 square inches—mothers, such as the one at the upper left, locate their own pups 83 percent of the time and so rarely nurse unrelated pups.

millions of others seemed impossible. Yet McCracken has shown, by comparing variations in blood enzymes between lactating females and suckling pups, that mothers successfully find and nurse their own young 83 percent of the time, apparently with the help of olfactory and auditory cues. Any nonparental suckling that takes place appears to be milk stealing on the part of an aggressive pup.

For species that form such enormous aggregations, such as the free-tailed bat, the benefits of creching (the pups stay warm, and the risk of predation to any one individual is reduced) outweigh the cost of occasionally nursing the wrong pup. Among free-tailed bats nonparental nursing—when it occurs—seems to be neither a form of kin selection nor reciprocal altruism but simply the result of random error.

For the past two summers my students and I have been studying nonparental nursing in evening bats in northern Missouri. These bats, like the free-tailed bats, form nursing colonies. Their colonies are relatively small, however, containing only from 30 to 200 adults, and are usually found in attics rather than in caves. Unlike the free-tailed bat, which gives birth to one pup per summer, the evening bat gives birth usually to two and sometimes to three pups at a time. Our studies indicate that a mother nurses her pups faithfully during the week following their birth but that as the pups age they tend to feed occasionally from other females. When a pup is about three weeks old, it is generally nursing from a female other than its mother about 20 percent of the time.

Is this a case of reciprocal altruism or kin selection? Because the females can be observed actively accepting or rejecting young pups who solicit milk from them, it seems they are discriminating between pups in some way. Preliminary evidence suggests that the females are selectively feeding relatives: analyses of blood-enzyme markers indicate that most often nonparental females are related—at least distantly—to the pups they nurse. In addition, data my graduate student J. Andrew Scherrer has recently collected indicate that each pup has a unique call and that calls made by relatives resemble each other. We suspect a nonparental female may recognize related pups by comparing their calls with those of her own pup.

Research on food sharing in bats illustrates a common theme in evolutionary biology: that similar behaviors seen in different species may result from entirely different evolutionary pressures. Although kin selection is widely regarded as a powerful and pervasive evolutionary force, under certain circumstances—such as whenever animals live in small groups and the potential for frequent aid giving among them is high—reciprocity is likely to be more beneficial than kin selection—provided that cheaters can be detected and excluded from the system.

Further understanding of the forces responsible for social evolution in vertebrates requires that the mechanisms underlying individual and kin recognition be identified. Our research on aid-giving behavior in bats demonstrates that the role both kin selection and reciprocity play in a society is dependent on the recognition capabilities of the animals in that society.

Determining the extent to which individuals recognize and interact preferentially with relatives should be greatly facilitated by modern molecular techniques (such as DNA fingerprinting), which significantly enhance the ability of investigators to measure relatedness among animals in the field. Because bats possess a sophisticated auditory system that enables them to navigate and capture prey, I believe careful study of their vocalizations in social situations may yield exciting information about the mechanisms by which animals recognize their relatives and close associates. The results of such study should in turn elucidate much about vertebrate social behavior in general.

FURTHER READING

THE EVOLUTION OF COOPERATION. Robert Axelrod. Basic Books, Inc., 1984.

SOCIAL EVOLUTION. Robert Trivers. Benjamin-Cummings Publishing Co., 1985.

THE NATURAL HISTORY OF VAMPIRE BATS. Edited by Arthur M. Greenhall and Uwe Schmidt. CRC Press, 1988.

RECIPROCAL ALTRUISM IN BATS AND OTHER MAMMALS. Gerald S. Wilkinson in *Ethology and Sociobiology*, Vol. 9, Nos. 2–4, pages 85–100; July, 1988.

Introducing...

CARE *of the* SURGICAL PATIENT

from SCIENTIFIC AMERICAN *Medicine*

Because the quality of your care depends on the quality of your information.

Treating pre and post operative patients poses a unique set of challenges. Yet in one way it's no different than any other practice issue.

Doing it well takes the right information.

That's why SCIENTIFIC AMERICAN *Medicine* is pleased to announce the publication of *CARE of the SURGICAL PATIENT.*

The definitive resource on pre and post-operative care.

CARE of the SURGICAL PATIENT gives you ready access to the most authoritative and current information on pre and post-operative standards available anywhere.

Written and designed by prominent surgeons under the supervision of the American College of Surgeons' Committee on Pre and Postoperative Care, *CARE of the SURGICAL PATIENT* provides two volumes – over 1,500 pages – of practical information on both critical and elective care.

And, *CARE of the SURGICAL PATIENT* is updated twice a year, with each surgeon-author reviewing his own specialty. Updates include new information on significant topics, such as current developments on AIDS.

In short, *CARE of the SURGICAL PATIENT* presents the standards for pre and postoperative treatment. You simply won't find a more important resource. Or one that organizes its information in such an intelligent way.

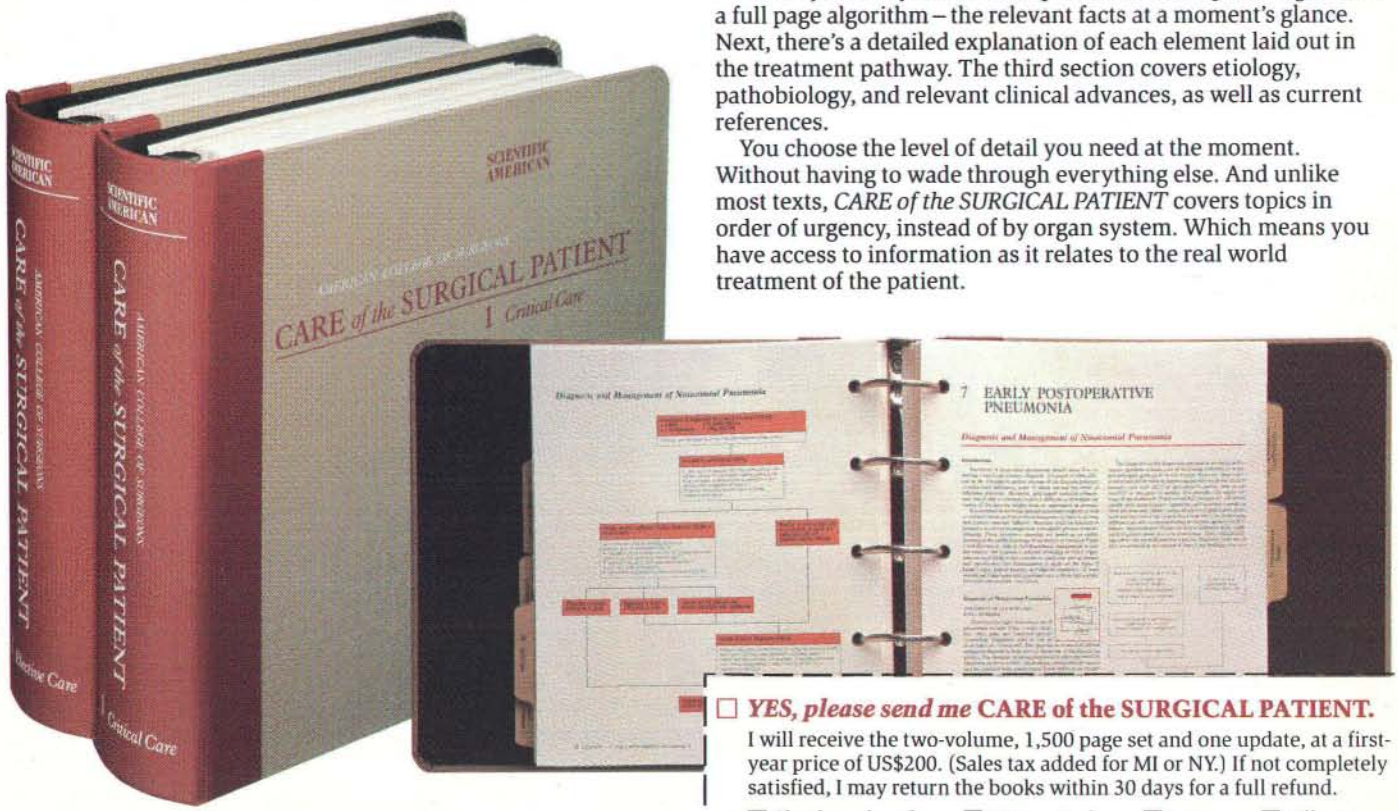
A unique system for rapid information retrieval.

CARE of the SURGICAL PATIENT is designed to get you the information you need, the way you need it.

Quickly. And intelligently.

The key is the system's three-part format. Chapters begin with a full page algorithm – the relevant facts at a moment's glance. Next, there's a detailed explanation of each element laid out in the treatment pathway. The third section covers etiology, pathobiology, and relevant clinical advances, as well as current references.

You choose the level of detail you need at the moment. Without having to wade through everything else. And unlike most texts, *CARE of the SURGICAL PATIENT* covers topics in order of urgency, instead of by organ system. Which means you have access to information as it relates to the real world treatment of the patient.



Try CARE of the SURGICAL PATIENT *Free for 30 days.*

You'll find it the most valuable resource on pre and post-operative care that's ever been published. And if you're not satisfied, just return it. No risk. No obligation.

CARE of the SURGICAL PATIENT, from SCIENTIFIC AMERICAN *Medicine*. No other resource helps you keep up better.

And the better you keep up, the better your care.

☐ ***YES, please send me*** **CARE of the SURGICAL PATIENT.**

I will receive the two-volume, 1,500 page set and one update, at a first-year price of US$200. (Sales tax added for MI or NY.) If not completely satisfied, I may return the books within 30 days for a full refund.

☐ Check enclosed ☐ MasterCard ☐ VISA ☐ Bill me

Acct.# _____ Exp. Date _____

Name _____

Address _____

_____

City _____ State _____ Zip _____

Specialty _____

Or call toll free 1-800 345 8112.

SCIENTIFIC AMERICAN MEDICINE 415 Madison Avenue, New York, NY 10017 9511

The Archaeology of Novgorod

Once the center of one of Europe's largest states, medieval Novgorod is nearly perfectly preserved. Among the finds are 700 birch-bark manuscripts that give an intimate view of the city's life

by Valentin L. Yanin

Any lover of classic films has watched Alexander Nevsky set forth on April 5, 1242, accompanied by the music of Prokofiev, to defeat the Teutonic Knights in the famous battle on the ice. Near the film's opening Alexander, prince of Novgorod, is summoned by the city's noblemen from his distant estate to defend Novgorod and free nearby Pskov from the invaders.

The fact that the prince does not come of his own accord but is summoned is a historical detail whose significance is probably lost on Western audiences. By the 13th century Novgorod was an "aristocratic republic" ruled by nobles, not princes. As is written of Nevsky's grandfather in the *Chronicle of Novgorod 1016-1471*: "The men of Novgorod showed Prince Vsevolod the road. 'We do not want thee, go whither thou wilt.'" Alexander had himself been expelled from the city for meddling in its affairs; when he was summoned to fight the Teutonic Knights, it was as a hired sword.

Novgorod was unique not only in its government but also in its religious tolerance and in its political and commercial power, which extended from Estonia in the west to beyond the Ural Mountains in the east and from the Arctic Ocean in the north to the upper Volga in the south. In short, it was one of Europe's largest states, maintaining trade and cultural links with Western Europe as well as central Asia and the Middle East. So powerful was the city that by the 14th century it had assumed the title "Lord Novgorod the Great." Although in 1478 it yielded its primacy to Moscow and eventually lost its economic importance, Novgorod has remained unusual in one other respect: the city was built on clay strata that have almost perfectly preserved its past, giving an unexcelled view into medieval Russian life.

VALENTIN L. YANIN is professor of history at Moscow M. V. Lomonosov State University and a corresponding member of the U.S.S.R. Academy of Sciences. He received a candidate degree from Moscow State in 1954, a doctorate in 1963 and became a professor there in the following year. Yanin has participated in the Novgorod excavations since 1947; he has directed them since 1962. As a second profession he collects early gramophone records.

Novgorod, or "New Town," was founded in the early ninth century, which makes it one of the oldest Russian cities, although its very name suggests that it was preceded by an even more ancient "Old Town." From the beginning Novgorod was the center of a vast territory settled by a variety of ethnic groups, including Slavs, Krivichi and Urgo-Finnish tribes. Traditionally the city was thought to have arisen as an outpost of Kiev, as Slavs moved northward from the Dnieper River to the Volkhov, which runs through Novgorod, and to adjacent Lake Ilmen. Recent findings suggest, however, that the Novgorodian Slavs probably arrived from Czechia and Poland. By the close of the ninth century Novgorod had merged with Kiev and enthroned a single prince, who moved his capital to the banks of the Dnieper. Among the evidence for the separate origins of Kiev and Novgorod is the existence (until the 13th century) of northern and southern monetary systems and two different systems of weights and measures.

As early as the 11th century Novgorod's accumulation of power led it to a prolonged struggle for independence from Kievan princes. It was during this struggle that Novgorod developed its peculiar form of government involving the institution of *posadnichestvo:* a popular assembly called the *veche* elected a wealthy boyar—a member of the highest-ranking aristocracy—to become *posadnik*. The *posadnik*'s function was basically to protect the boyars against the prince and his cohort. On several occasions the Novgorodian boyars refused altogether to accept the prince appointed by Kiev. (In one case they relented—on the condition the designee have two heads.) The struggle culminated in 1136 with the famous explusion of Vsevolod from the city.

Although this "revolution" did not eliminate the rule of the princes altogether, it did severely circumscribe their authority; the prince became essentially a hired official. Princely power was already nonhereditary in Novgorod, and once having won the right to invite and expel princes, Novgorodians transformed the throne into a symbol of political union with those principalities from which they invited princes. Such alliances proved mutually advantageous: for centuries Novgorod protected northwest Russia from invaders, and the allied principalities enhanced Novgorod's stature.

Just as striking as Novgorod's expulsion of the princes in 1136 was its seizure in 1156 from Kiev of the right to elect its own archbishop. The institution of these elections, which were carried out through the *veche,* gave Novgorod a virtually independent religious administration. With these actions Novgorod entered its golden age. While the remainder of Rus, as Russia

KREMLIN OF NOVGOROD as it exists today is seen in this aerial view. According to the chronicles, construction of the Kremlin was begun in 1044. St. Sophia's Cathedral, in the center, was built between 1045 and 1050. In 1116 the citadel was enlarged to its present size; it underwent major reconstruction after the city surrendered to Moscow in 1478. The carillion to the left dates from 1439 and the tall bell tower to the right from 1673. The long building in the upper part of the Kremlin is the Novgorod Museum of History, Architecture and Art; it houses some of the 130,000 artifacts excavated over the past 60 years.

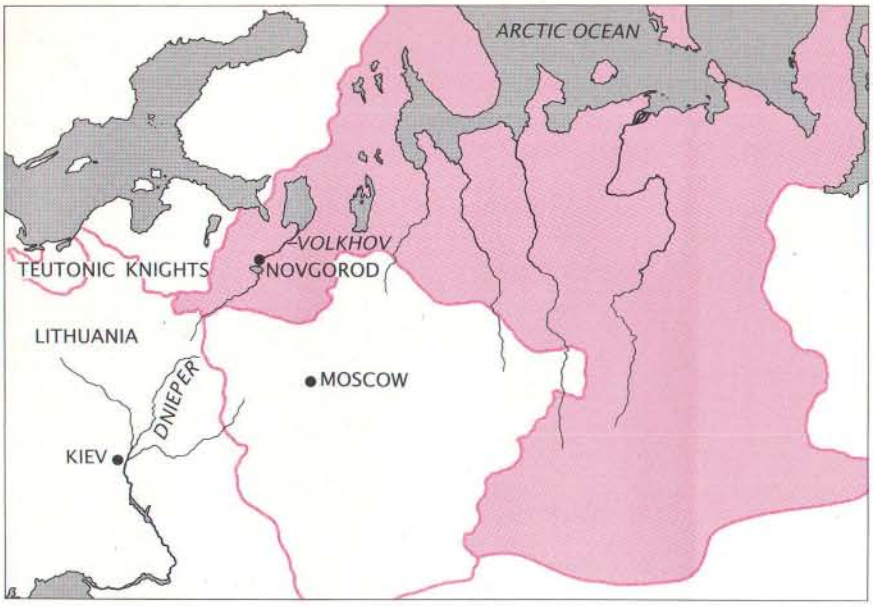
was then called, was subjected to the Tartar-Mongol conquest, Novgorod's willingness to buy off the invaders with tribute, the city's distance from the main body of the horde and its marshy surroundings all contributed to its autonomy, which continued until it surrendered to Moscow in 1478. That year marks the creation of the Russian state, and from then on Novgorod's story is one of decline: trade routes shifted, the Swedes overran the city between 1611 and 1617 and by the late 18th century it had been rebuilt in a way that bore little resemblance to the original. Today Novgorod is a modern Russian city several hours from Leningrad with a population of 200,000.

Medieval Novgorod has not been lost, however. Excavations begun there in 1929 under Artemii Vladimirovich Artsikhovsky of Moscow M. V. Lomonosov State University have brought to light more than 130,000 artifacts in a remarkable state of preservation, including one of the great archaeological finds of the 20th century: 700 birch-bark manuscripts that, by giving an intimate view of daily cares, bring the city to life in a way that the standard textbook history is unable to do.

Ancient Novgorod owes its nearly perfect state of preservation to the fact that Novgorod is built on compacted clay strata, which prevent the drainage of floodwater and rainwater. The soil above the clay remains saturated with moisture, which has preserved all organic remains—from the first human settlements onward—intact. The high moisture content prevented city dwellers from digging deep foundations or vaults and also forced them to pave their streets with wooden planks. At the same time, Novgorod's active life led to a rapid growth of the cultural layer, the layer of deposits left by human activity. The cultural layer's growth of one centimeter a year required that every two or three decades a new pavement of wooden planks be laid over the previous pavement. As a result, beneath modern Novgorod exists a multitiered structure of wooden streets; in some places more than 30 tiers reach a thickness of nearly 10 meters and stretch back 10 centuries.

Along with the tiers of pavements are countless houses or, more accurately, those parts of houses (usually the lower rows of logs) that survived the frequent fires that destroyed many medieval Russian cities. With the remains of houses, of course, come many household items. The fact that the household remains can be linked to a given layer of street paving has facilitated the physical placement of virtually all the finds and helped to establish their relative ages.

Because of the abundance of old wooden structures, one can accurately date the Novgorod finds by the dendrochronological, or tree-ring, method. As originally developed by Andrew E. Douglass in the U.S. at the beginning of the century, the technique relied on such long-lived species as the sequoia or yellow pine, which do not grow in Eastern Europe. There the traditional construction material was ordinary pine, whose life span rarely exceeds 150 years. Boris A. Kolchin of the Institute of Archaeology of the U.S.S.R. Academy of Sciences was nonetheless able to match the ring pattern of suc-



GREAT NOVGOROD ruled lands that in the 14th and 15th centuries stretched from the Arctic Ocean in the north to the Russian principalities in the south; from beyond the Ural Mountains in the east to lands occupied by the Teutonic Knights in the west. The vast holdings made Novgorod one of the largest states in Europe.

cessive generations of trees, from medieval pines to living trees, in order to establish an absolute chronological scale. The dendrochronological approach is extremely accurate: an excavated log can be dated to within one year, which means that the age of every wooden structure in Novgorod is firmly established. Household items found between tiers can be dated to within 15 to 25 years.

Apart from accurate dating, another exceptional circumstance has contributed to the success of the Novgorod excavations. This has been the ability to excavate a relatively wide area—entire estates, for example, and even large parts of the old city. Such a circumstance is remarkable because Novgorod is not a dead city but one that continues to lead an active life. The wide-scale excavations have been made possible by a close cooperation between the Novgorod City Council and archaeologists. In the late 1960's an ordinance was passed that prohibited any construction until the prospective building site had been studied by archaeologists. Of particular importance has been the information obtained from the hundreds of bore holes that builders must make to determine the strength of projected foundations. From each bore hole investigators can read off the cultural layer's thickness at that location; combining information from all the bore holes makes it possible to map the size of the city at each stage of its growth and target the most interesting sites for excavation.

These favorable conditions have resulted in a treasure trove of artifacts: tools, household utensils and vessels, the fighting gear of both a foot soldier and a mounted warrior, parts of ships and their block-and-tackle mechanisms, jewelry, furniture and ornate wood carvings. Even clothes, leather slippers and musical instruments have been preserved in such a state that one would think with a little repair they could be worn or played. The sheer quantity of leather and wood objects is itself an important corrective to common perceptions. How misleading are museum exhibits that feature objects of metal, stone and glass, when 90 percent of household objects in medieval times were made of wood!

The excavations refute the so-called trading theory, which long dominated the historical accounts of Russian cities and which argued that Russian cities lacked skilled craftsmen and were forced to import finished items from abroad in return for produce. Although only 2 percent of old Novgorod has been studied, the digs have uncovered about 140 artisan workshops specializing in the manufacture of locks, leather products, jewelry, shoes, metal casting and so on. Breweries, bakeries and dyeing, weaving and glass shops have also been found. The items from all of these shops reveal a standard of workmanship and degree of specialization on a par with those of the most celebrated medieval centers of Western Europe and the Middle East.

Most of these workshops did not belong to free artisans but were attached to large estates belonging to wealthy boyars. An examination of knife manufacture in the 11th and 12th centuries serves to show how the "mass production" of items affected trade patterns and the activities of artisans, merchants and the boyars.

In 11th-century Novgorod artisans fashioned knives by the "packet technique," in which plates of softer iron were welded to both sides of a steel blade. The knife became a self-sharpening instrument: gradual wear of the outer plates revealed more and more of the steel blade, which could be used until it was totally abraded. By the first quarter of the 12th century the packet technique had been replaced by a simpler construction: a narrow strip of steel was welded to an iron base to form a thin cutting edge. Such simplification of manufacturing technique to meet broader market demand enabled the artisan to manufacture more articles faster, but of course the articles became less durable—a situation with which we are familiar today.

The sharply increased output required more imported raw materials and, consequently, more exports to pay for them. The exports tended to be furs, wax, honey, flax and valuable varieties of fish. Not surprisingly, the enhanced trading was accompanied by an expansion of the boyars' colonizing activities in northern territories such as those around Lake Onega and along the northern Dvina River.

A study of Novgorod's imports reveals the extent of its trading links and supports the picture of a thriving commercial empire. From the excavations it is apparent that Novgorod's imports consisted mainly of raw materials unavailable in the city. Nonferrous metals (gold, silver, copper, lead and tin) came from England, Sweden, Poland and Hungary; semiprecious stones came from the Urals and (via the Volga waterway) from Iran. Most of the many hair combs unearthed are of box-tree wood from the Caspian Sea region. Amber from the Baltic was the favorite material for beads and rings.

Metallographic and spectrographic studies provide more information. For instance, in the 14th-century layer a silver ingot weighing 150 kilograms has been unearthed. It bears the seals of the Polish king Kazimir the Great,

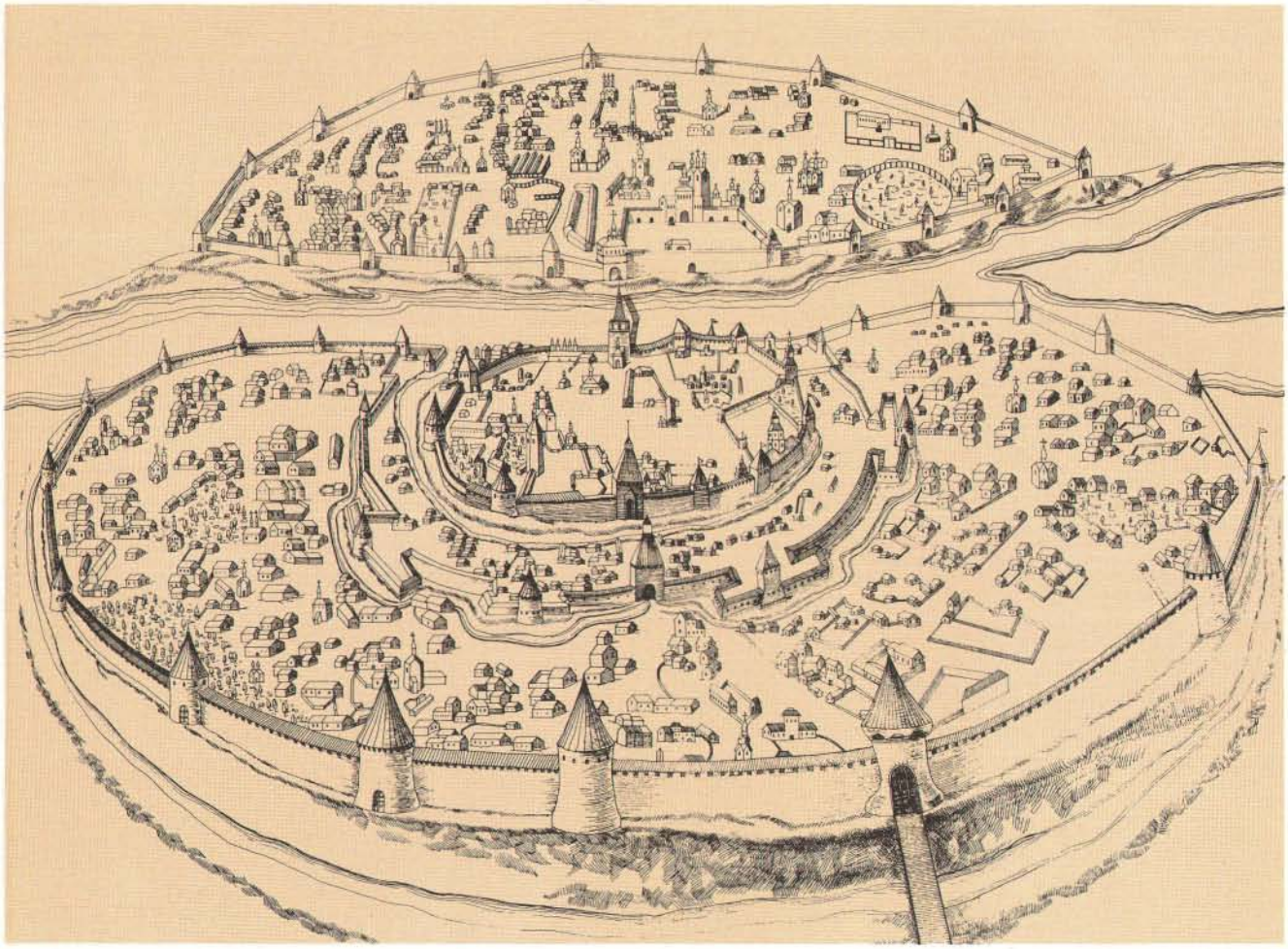
and analysis of the metal indicates that it originated near Kraków, in Poland. Similarly, rings from the 14th century typically contain two crystal segments held together by a glue that lends extra sparkle to the stone. Study of the crystals shows that they came from Madagascar; the shellac for the glue was from the Maldive Islands in the Indian Ocean. Needless to say, Novgorod had no direct trade links with these places. The materials reached Novgorod by way of the Western European cities with which it maintained stable ties.

Given the perfect state of preservation of Novgorod's wooden streets, houses, manors and artifacts, it is almost possible to imagine rich merchants concluding contracts with Swedes and Karelians, peasants selling wares at the market, innumerable sailing ships coming in along the Volkhov laden with expensive foreign wine, the smell of fish, shoemakers stitching sandals, women adorning themselves with furs and jewelry, street musicians plucking their psalteries. The only thing missing from Novgorod's past are the voices of its inhabitants.

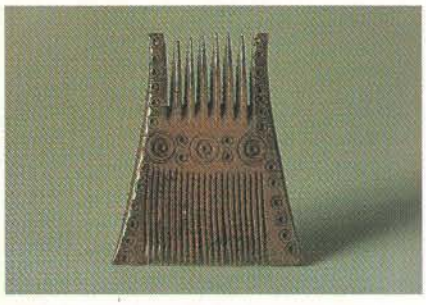
But on July 26, 1951, they too were provided. Excavating in the sixth tier (1369–1409) of ancient Kholop'ya ("Serfs") Street, archaeologists uncovered a birch-bark manuscript, the first of some 700 that have been discovered by now. These birch-bark letters, or *beresty,* date from the mid-11th century to the early 15th century and have also been found in several other old Russian cities, including Smolensk and Pskov.

The *beresty* of Novgorod vary widely in content. They include household information, strategic and political reports, lawsuits, peasants' grievances, technical instructions from craftsmen and orders for icons: "Greetings from the priest to Grechin. Paint me two six-winged seraphims on two icons for the top of the iconostasis. I kiss you. God will reward you or we'll make a deal." They include riddles: "There is a city between heaven and earth and to it comes an emissary not on a road, carrying an epistle that is not written." They include love letters: "From Nikita to Ulyanitsa. Marry me. I want you and you me. And as witness will be Ignato." They include practical problems: "From Boris to Nastasya. As soon as you receive this letter send me a man on horseback, since I have a lot to do here. Oh yes, send a shirt. I forgot one." And they include more universal matters: "Greetings from Nastasya to my brothers. My Boris is no longer among the living. Please, Lord, preserve me and my children."

With the discovery of the *beresty* it is no longer necessary to imagine the



NOVGOROD is depicted in an icon of the late 17th century. The layout closely resembles the city's plan of the 14th and 15th centuries, when old Novgorod reached its maximum size. In the 16th century the city lost its previous importance, and it expanded no further; these boundaries remained until a significant replanning was begun in 1788. In the foreground is the St. Sophia side, which has been most extensively excavated in the areas just above and below the Kremlin. In the past decade, many areas of the trading side, across the river, have also been excavated, allowing comparison of all parts of Novgorod.



ARTIFACTS excavated at Novgorod number about 130,000. Shown here from top to bottom are stone beads, a bone comb, a pagan amulet (exact replicas of which have been found near the Urals), a medallion depicting St. George on horseback and the reverse side of the medallion depicting intertwined snakes. The snakes are pagan symbols; the existence of many such artifacts suggests that paganism and Christianity coexisted in Novgorod until the 16th century.

thoughts of the inhabitants of medieval Novgorod; the inhabitants speak for themselves. Their joys and sorrows, conflicts, friends and enemies are all before us, alive on the streets of Novgorod. How were we so lucky to find them? It should first be mentioned that the *beresty* are not ordinary pieces of birch bark; they were specially prepared by being boiled in water to soften the skin, after which the coarser layers were removed. The *beresty* were thus not occasional "accidents" but letters. They performed all the functions of today's letters and were discarded as such; that is why they are fairly abundant.

In addition to shedding light on the daily lives of medieval Novgorodians, the *beresty* resolved some mysteries relating to artistic masterpieces. In the remains of a building destroyed by fire in 1207, archaeologists hit on a cache of paints, imported mineral dyes and other material for the preparation of paints and varnishes. They also uncovered small wooden panels and traces of bronze frames, all of which meant the building was an icon studio. Indeed, more than a dozen *beresty* have been found that place orders for icons, five of them naming as the painter one Olisei-Grechin. According to the chronicles, in 1196 Olisei-Grechin painted the frescoes in the Church of the Virgin in the Novgorod Kremlin. Unfortunately, neither the church nor the frescoes have survived.

His great reputation, however, led art historians to speculate that Olisei-Grechin headed the team of from eight to 10 artists that painted the most famous ensemble of medieval Russian frescoes: those in the Church of Our Savior on Nereditsa Hill in the outskirts of Novgorod. Nevertheless, the frescoes were not signed, and the historians had despaired of ever identifying the anonymous master. With the discovery of the *beresty*, the identification was clinched: the grammatical mistakes that appear in Olisei-Grechin's letters are also found on the inscriptions below the frescoes. The two artists are one and the same.

The importance of the *beresty* is not confined to "local" issues; the birch-bark manuscripts also touch on larger questions connected with the city, such as illiteracy, the emergence of the republic and the functioning of the *veche*. Before 1951 it was widely thought that the people of ancient Rus were overwhelmingly illiterate. With the continued unearthing of *beresty*, from all levels of society, it becomes more difficult to make this argument, especially in light of the large number of school exercises that have been found.

Other preconceptions have had to be rescrutinized. In the 19th and early 20th centuries historians thought most of the power in Novgorod resided with the rich merchants; the indelible connection between Novgorod and merchants goes back at least to the 12th century when the *byliny*, or heroic tales, sang of the adventures of the rich merchant Sadko (later the hero of Rimsky-Korsakov's opera *Sadko*). In fact, Novgorod was often referred to as a "commercial republic."

The *beresty* give a somewhat different picture. Dozens of letters show that power in Novgorod belonged not so much to the merchant class as to the boyar aristocracy. The 1136 uprising against Prince Vsevolod marked the triumph of this aristocracy, which had already managed to accumulate considerable power by the institution of a patrimonial, or hereditary, estate system.

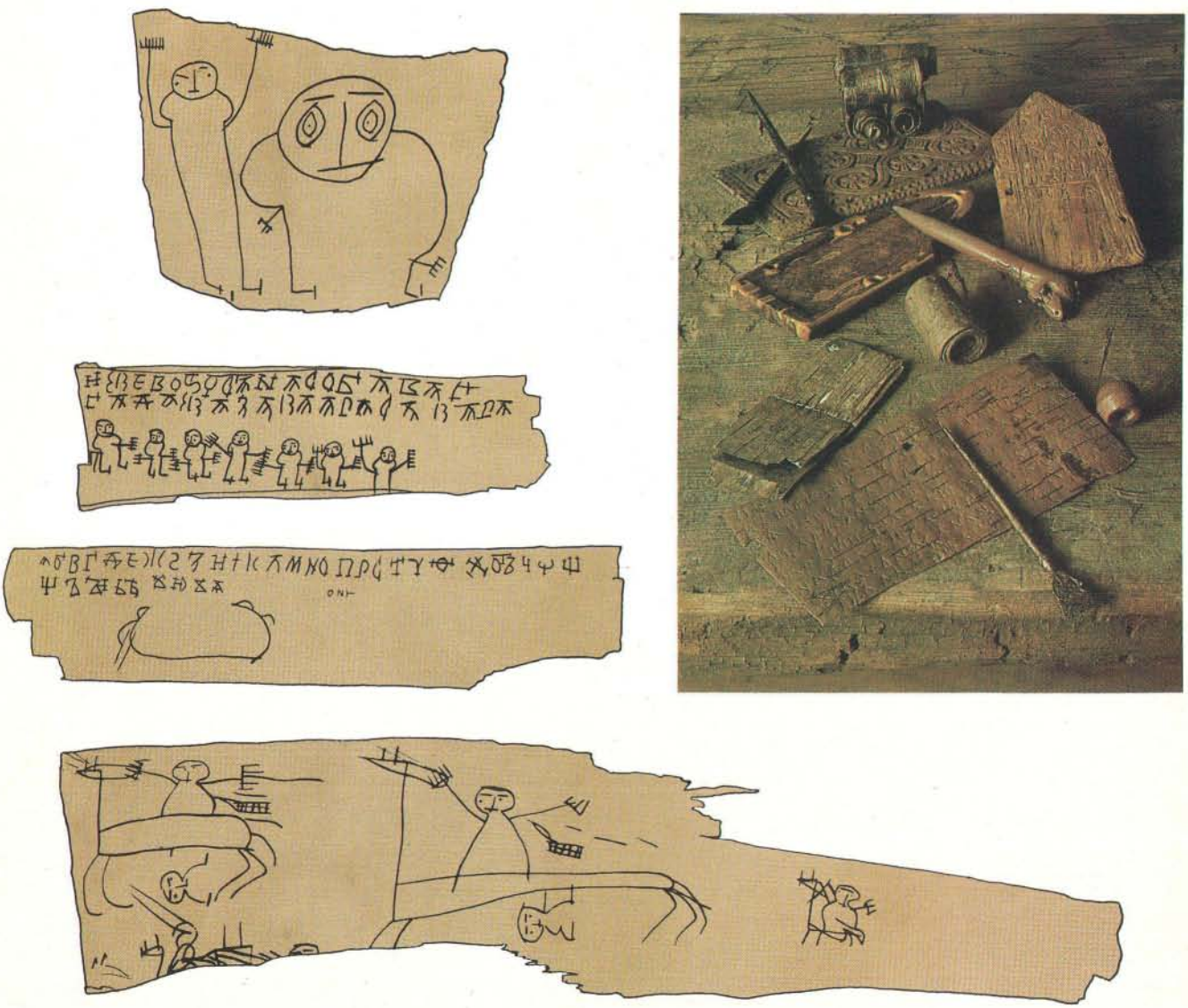
Prominent boyar families, for instance, apparently owned large sections of city territory; a clan of families might own as many as 15 manorial estates, each of 1,200 to 1,500 square meters. The system also assured the establishment of a closed, self-reliant economy, because practically every boyar estate had workshops and, through the clan, every family had access to a wide range of goods and services. After internal consumption, any surplus was sold in the city market, thus firmly linking the estate owners with the larger Novgorod economy. Although such a system consolidated the boyars as a social class, the existence of the clans prevented artisans of Novgorod from organizing into trade guilds.

In light of this rather severe state of affairs, to what extent can the political institutions of Novgorod be termed democratic? The question is not new; it first arose nearly two centuries ago when the *veche* system was held up as one of the cradles of self-government and independence. The *veche* has traditionally been visualized as a huge assembly of thousands of city dwellers, merchants and boyars, all standing and voicing their opinions in the best tradition of democracy and openness. The 18th-century Russian radical Alexander Radischev wrote that, "Novgorod had a popular government.... The people assembled in the veche were its true sovereign." The poet Mikhail Y. Lermontov's view was even more romantic:

Hail, sacred cradle of warrior Slavs!
Arrived from foreign lands, I gaze
With rapture at the gloomy walls
Through which the centuries of change
Passed harmlessly; where the veche bell
Alone did serve the cause of freedom...

The archaeological finds at Novgorod reveal a somewhat more sober picture. Chronicles and official documents from medieval Novgorod indicate that the *veche* normally assembled at the *Yaroslavovo dvorishche,* the "big square around Yaroslav's palace." Not surprisingly, the search for wooden planks from the *veche* square—or at least from some large open space near the palace that was free of any structures—became a primary goal of the Novgorod digs. But numerous excavations in almost every direction found no such open space: remains of homes and outbuildings were everywhere.

At the present time the only area left to be searched adjoins the western side of the Nikolsky Cathedral, which dates from 1113. We cannot be absolutely certain that this was the location of the *veche,* because the area is currently occupied by stone buildings of a later period. Yet it is the last available candidate, and certain written sources from no later than the beginning of the 13th century state that the *veche* used to assemble at "St. Nikola's." If this is the correct site, one thing is already apparent: the *veche* was not a huge assembly. The



BIRCH-BARK manuscripts, or *beresty,* are the most significant finds of the Novgorod digs. Shown in the photograph are a *beresta* of the 14th century (*foreground*), a birch-bark book (*middle*) and metal and bone styli for writing on the bark. Also shown near the top are two wooden writing tablets for students, which were meant to be covered with wax. The holes in the first tablet served as "waxwells." The wide ends of the styli served to erase the wax text. The reverse side of the second tablet (*upper right*), dating from the 14th century, is carved with an alphabet, presumably for use in school. To the left and below the photograph are shown actual school exercises written on birch bark by the boy Onfim in the first half of the 13th century. On 10 pages are the letters of the alphabet in a form of Old Slavic (*a, b, v, g...*) and lessons on the formation of syllables (*ba, va, ga, da, be, ve, ge, de...*). Onfim covered the remaining space with drawings of himself, battle scenes and pictures of his teacher. The draftsmanship suggests that Onfim was six or seven years old at the time. Evidently, students did their first lessons on wax tablets of the type shown; they then advanced to birch bark, which required a firm hand.

area next to St. Nikola's is no larger than 2,000 square meters, and the chronicles indicate that the participants were seated during the meetings, suggesting that the square was probably covered with a wooden platform and benches. Under those conditions the site could have accommodated only 300 to 500 people. The estimate is supported by a German source of 1331, which refers to Novgorod's governing body as the "300 golden belts." Finally, about 400 manorial estates of boyars and other wealthy people have so far been counted.

The *veche,* then, was an assembly of the wealthiest townsmen, primarily boyars, whose power was based on vast landholdings. Its sole claim to democracy rested on the public nature of its proceedings. Rather than caucus behind closed doors, the *veche* held its deliberations in full view of the public, which could, by shouts of acclamation or censure, deceive itself into thinking it was participating in the decision-making process.

Although the digs have helped illuminate the nature of Novgorod's republic, they have also presented researchers with the difficult task of explaining the origin of boyar statehood. For a long time it was assumed that, with the establishment of Kiev as the capital of Rus in the late ninth century, the Novgorodian princes directed their energy southward and showed only a passing interest in Novgorod itself. The inattention presumably set the stage for a gradual takeover by the boyars of property that is known to have been held communally before then.

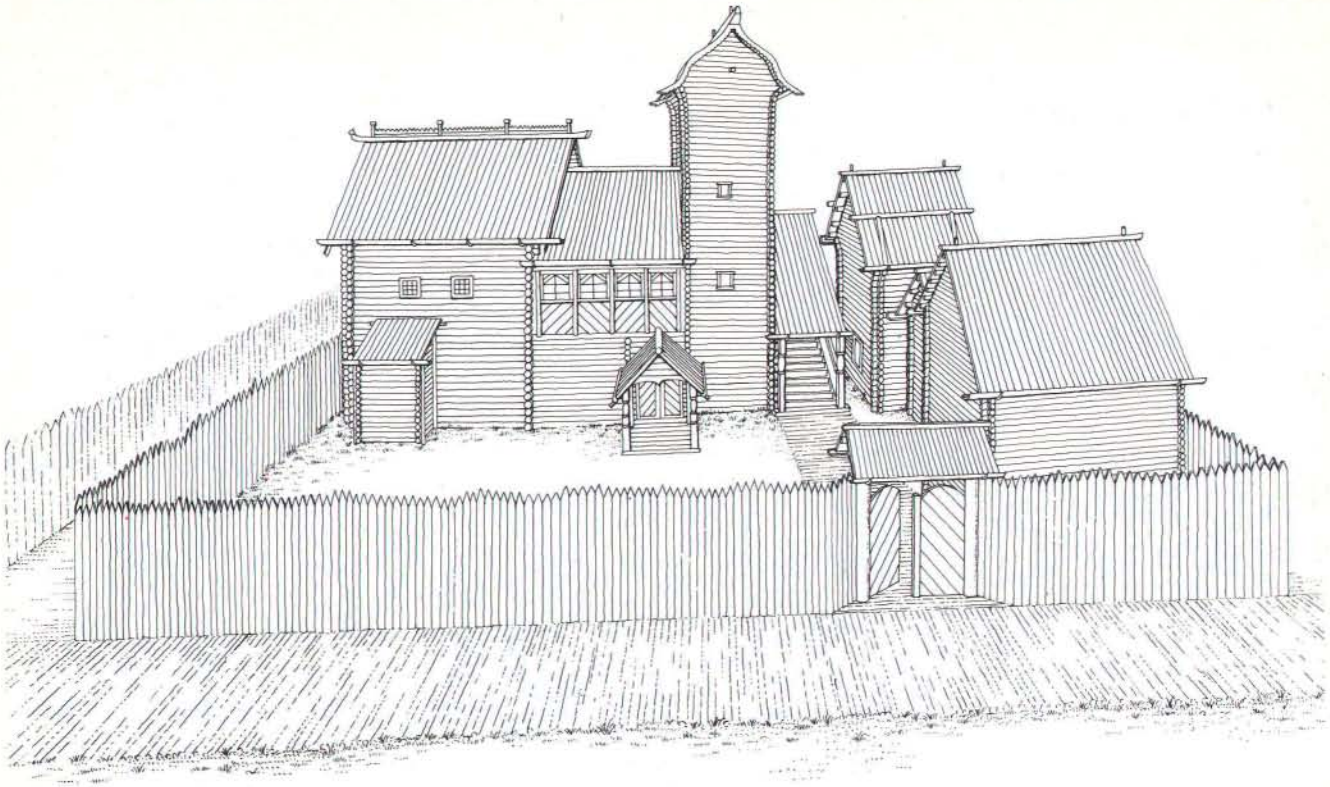
Recent studies of a number of sources, including *beresty* of the 11th and 12th centuries, suggest that the boyar patrimonial estate system originated only in the first quarter of the 12th century, by which time the aristocrats had already gained so much power they were on the verge of expelling the prince. Therefore, the ascendancy of the boyars, which had been assumed to be the effect of the patrimonial system, instead proved to be the cause. Political gains made by the boyars in their struggle with the Novgorodian princes created the patrimonial estate system, which did not reach the peak of its development until the 14th century.

The rise of the boyar class has left certain traces. For instance, nine wooden cylinders marked with *tamgas,* or emblems of princely authority, have been found in excavations dating from the 10th and 11th centuries—before the boyars possessed patrimonial estates. These cylinders, which are carved in an unusual fashion, went unexplained for 30 years. With the discovery of certain inscriptions, such as "the tax collector's three big silver coins" and "the sack of Polotvitsa the tax collector is secured with this seal," it has become clear that the cylinders are seals for money bags. The fact that they have been found at manorial estates, and not just at the prince's court, indicates that the early boyars participated in the collection of taxes. According to the Russian Pravda, the oldest written law of ancient Rus, the bulk of the taxes went to the prince, a tenth went to the church and a fixed share went to the tax collector. In this way the boyars managed to enhance their own income.

Access to the circulation of money in turn resulted in the widespread practice of moneylending. *Beresta* number 526, dating from about 1080, reveals a web of moneylending activity that is truly staggering; it encompasses essentially all the Novgorodian lands of the time. Although the conclusion is debatable, it is reasonable to assume that by such activities the boyars gradually increased their economic power until they were in a position to oust the prince and relegate him to a position inferior to that of the *posadnik.*

The most difficult problem posed by the excavations is that of reconciling Novgorod's unusual form of government and laws with the "Normanist" view of how these institutions evolved. Normanists assert that the groundwork for Russia's government and law was laid down by the Norsemen, or Normans, who were led by Prince Rurik and supposedly were invited in 862 by the quarreling Slavs to rule over them. According to the *Russian Primary Chronicle,* which dates from the 12th century, Rurik actually settled in the vicinity of Novgorod. The accuracy of the *Chronicle* is constantly being disputed; nevertheless, the antiquity of Novgorod's

ANNA writes to her brother Klimyata in the late 12th century to inform him of a legal suit that has been brought against her and her daughter. A certain Kostyantin has accused them of lending without authorization money that he had entrusted to them and then keeping the interest for themselves. "Brother," Anna writes, "say to him in front of witnesses, 'Why were you angry at my sister and her daughter? Why did you call them whores?'" Anna goes on to say that her husband, having heard about the accusations, threatened to kill her. She declares: "If they find any witnesses to the accusations brought against me, then I am not a sister to you or a wife to my husband. You can then kill me yourself, but neither I nor my daughter is guilty of anything." Kostyantin has threatened Anna with a fine, which he would serve in the administrative center of Anna's village. Anna herself, however, intends to bring suit against Kostyantin for disgracing her womanly honor. Until this letter was discovered, details of legal procedure were known only from the 15th-century Judicial Code of Pskov, which led scholars to suppose a late origin for this important part of the Russian legal system. Anna's letter, containing details fully analogous to the Pskov Code, provides evidence that the same legal processes existed three centuries earlier than had been thought.



DWELLINGS OF WOOD were universal in medieval Russia. The building on the left is an artist's reconstruction of the Novgorod home of Olisei-Grechin, who is considered one of the greatest icon painters of the 12th century. The reconstruction is based on the actual excavation of the house and on architectural motifs found on frescoes in the church on Neriditsa Hill.

governmental institutions would suggest that the Normans were not given free rein but were invited in by an already well-organized group and granted strictly prescribed powers. The *veche* in particular had its roots in earlier social institutions, which proved more durable than princely power imposed from the outside.

Similar questions arise about Russian law. It has often been maintained that Russian law appeared rather late, in the second half of the 11th century with the Russian Pravda and in the 15th century with the Judicial Code of Pskov. The main points of law encoded in these documents were thought to have been formulated after Rus conversion to Christianity in 988. For instance, the Russian Pravda bans blood feuds in favor of a system of fines levied by the state. Among the Novgorod finds, however, is *beresta* number 531, which dates from the late 12th century and describes laws that were thought to have come into effect only three centuries later, with the Code of Pskov [*see illustration on opposite page*]. Moreover, several of the tax-collection cylinders described above bear inscriptions showing that the system of blood-feud fines detailed by the Russian Pravda was in effect as early as 970. Consequently, the Novgorod finds strongly suggest that Russian law came into existence at least three centuries—and perhaps five centuries—earlier than previously thought and that it did not undergo any fundamental change in the course of time.

The archaeological excavations at Novgorod touch on a variety of issues, from the daily lives of its inhabitants to broader questions of ancient Russian government and law. What is more, the successes to date have come from digs that cover only 2 percent of the old city's surface. It is quite possible that the most important finds have yet to be made. For instance, we have estimated that no less than 20,000 *beresty* are waiting to be exhumed. One can only envy the future historians who will read these documents and delve deeper into the almost undisturbed privacy of medieval Novgorod.

In a larger context, it is difficult to overestimate the importance of Novgorod in medieval Russia. Fully half of the Russian books surviving from that period have been found there. The city's place in Europe as a whole was also extremely significant. Even now it is known that a special corporation in northern Germany existed to trade with Novgorod, that Hanseatic, Dutch and German merchants lived in Novgorod and that a Russian merchant community existed in Holland. In a numismatic duplication unique in Europe, both Novgorodian and Venetian coins feature the city's leader receiving the symbols of state from each city's respective patron saint. Is this a coincidence or an example of borrowing from a distant government? The answer is not known. But future discoveries may further clarify the relations of the European powers with the city that was once justly called Lord Novgorod the Great.

FURTHER READING

V. L. YANIN AND THE HISTORY OF NOVGOROD. V. L. Yanin in *Slavic Review*, Vol. 33, No. 1, page 114; March, 1974.

THE ESTATE OF A NOVGOROD ARTIST OF THE 12TH CENTURY. B. A. Kolchin, A. C. Khoroshev and V. L. Yanin. Nauka, Moscow, 1981.

THE NOVGORODIAN FEUDAL PATRIMONIAL ESTATE. V. L. Yanin. Nauka, Moscow, 1981.

50 YEARS OF THE NOVGOROD EXCAVATIONS. Edited by B. A. Kolchin and V. L. Yanin. Nauka, Moscow, 1982.

THE LANGUAGE OF THE BIRCH-BARK MANUSCRIPTS. V. L. Yanin and A. A. Zaliznyak in *The Future of Science*, No. 20, pages 256–271; 1987.

Positive Feedbacks in the Economy

A new economic theory elucidates mechanisms whereby small chance events early in the history of an industry or technology can tilt the competitive balance

by W. Brian Arthur

Conventional economic theory is built on the assumption of diminishing returns. Economic actions engender a negative feedback that leads to a predictable equilibrium for prices and market shares. Such feedback tends to stabilize the economy because any major changes will be offset by the very reactions they generate. The high oil prices of the 1970's encouraged energy conservation and increased oil exploration, precipitating a predictable drop in prices by the early 1980's. According to conventional theory, the equilibrium marks the "best" outcome possible under the circumstances: the most efficient use and allocation of resources.

Such an agreeable picture often does violence to reality. In many parts of the economy, stabilizing forces appear not to operate. Instead positive feedback magnifies the effects of small economic shifts; the economic models that describe such effects differ vastly from the conventional ones. Diminishing returns imply a single equilibrium point for the economy, but positive feedback—increasing returns—makes for many possible equilibrium points. There is no guarantee that the particular economic outcome selected from among the many alternatives will be the "best" one. Furthermore, once random economic events select a particular path, the choice may become locked-in regardless of the advantages of the alternatives. If one product or nation in a competitive marketplace gets ahead by "chance," it tends to stay ahead and even increase its lead. Predictable, shared markets are no longer guaranteed.

During the past few years I and other economic theorists at Stanford University, the Santa Fe Institute in New Mexico and elsewhere have been developing a view of the economy based on positive feedback. Increasing-returns economics has roots that go back 70 years or more, but its application to the economy as a whole is largely new. The theory has strong parallels with modern nonlinear physics (instead of the pre-20th-century physical models that underlie conventional economics), it requires new and challenging mathematical techniques and it appears to be the appropriate theory for understanding modern high-technology economies.

The history of the videocassette recorder furnishes a simple example of positive feedback. The VCR market started out with two competing formats selling at about the same price: VHS and Beta. Each format could realize increasing returns as its market share increased: large numbers of VHS recorders would encourage video outlets to stock more prerecorded tapes in VHS format, thereby enhancing the value of owning a VHS recorder and leading more people to buy one. (The same would, of course, be true for Beta-format players.) In this way, a small gain in market share would improve the competitive position of one system and help it further increase its lead.

Such a market is initially unstable. Both systems were introduced at about the same time and so began with roughly equal market shares; those shares fluctuated early on because of external circumstance, "luck" and corporate maneuvering. Increasing returns on early gains eventually tilted the competition toward VHS: it accumulated enough of an advantage to take virtually the entire VCR market. Yet it would have been impossible at the outset of the competition to say which system would win, which of the two possible equilibria would be selected. Furthermore, if the claim that Beta was technically superior is true, then the market's choice did not represent the best economic outcome.

Conventional economic theory offers a different view of competition between two technologies or products performing the same function. An example is the competition between water and coal to generate electricity. As hydroelectric plants take more of the market, engineers must exploit more costly dam sites, thereby increasing the chance that a coal-fired plant will be cheaper. As coal plants take more of the market, they bid up the price of coal (or trigger the imposition of costly pollution controls) and so tip the balance toward hydropower. The two technologies end up sharing the market in a predictable proportion that best exploits the potentials of each, in contrast to what happened to the two video-recorder systems.

The evolution of the VCR market would not have surprised the great Victorian economist Alfred Marshall, one of the founders of today's conventional economics. In his 1890 *Principles of Economics,* he noted that if firms' production costs fall as their market shares increase, a firm that simply by good fortune gained a high

W. BRIAN ARTHUR is Morrison Professor of Population Studies and Economics at Stanford University. He obtained his Ph.D. from the University of California, Berkeley, in 1973 and holds graduate degrees in operations research, economics and mathematics. Until recently Arthur was on leave at the Santa Fe Institute, a research institute dedicated to the study of complex systems. There he directed a team of economists, physicists, biologists and others investigating behavior of the economy as an evolving, complex system.

proportion of the market early on would be able to best its rivals; "whatever firm first gets a good start" would corner the market. Marshall did not follow up this observation, however, and theoretical economics has until recently largely ignored it.

Marshall did not believe that increasing returns applied everywhere; agriculture and mining—the mainstays of the economies of his time—were subject to diminishing returns caused by limited amounts of fertile land or high-quality ore deposits. Manufacturing, on the other hand, enjoyed increasing returns because large plants allowed improved organization. Modern economists do not see economies of scale as a reliable source of increasing returns. Sometimes large plants have proved more economical; often they have not.

I would update Marshall's insight by observing that the parts of the economy that are resource-based (agriculture, bulk-goods production, mining) are still for the most part subject to diminishing returns. Here conventional economics rightly holds sway. The parts of the economy that are knowledge-based, on the other hand, are largely subject to increasing returns. Products such as computers, pharmaceuticals, missiles, aircraft, automobiles, software, telecommunications equipment or fiber optics are complicated to design and to manufacture. They require large initial investments in research, development and tooling, but once sales begin, incremental production is relatively cheap. A new airframe or aircraft engine, for example, typically costs between $2 and $3 billion to design, develop, certify and put into production. Each copy thereafter costs perhaps $50 to $100 million. As more units are built, unit costs continue to fall and profits increase.

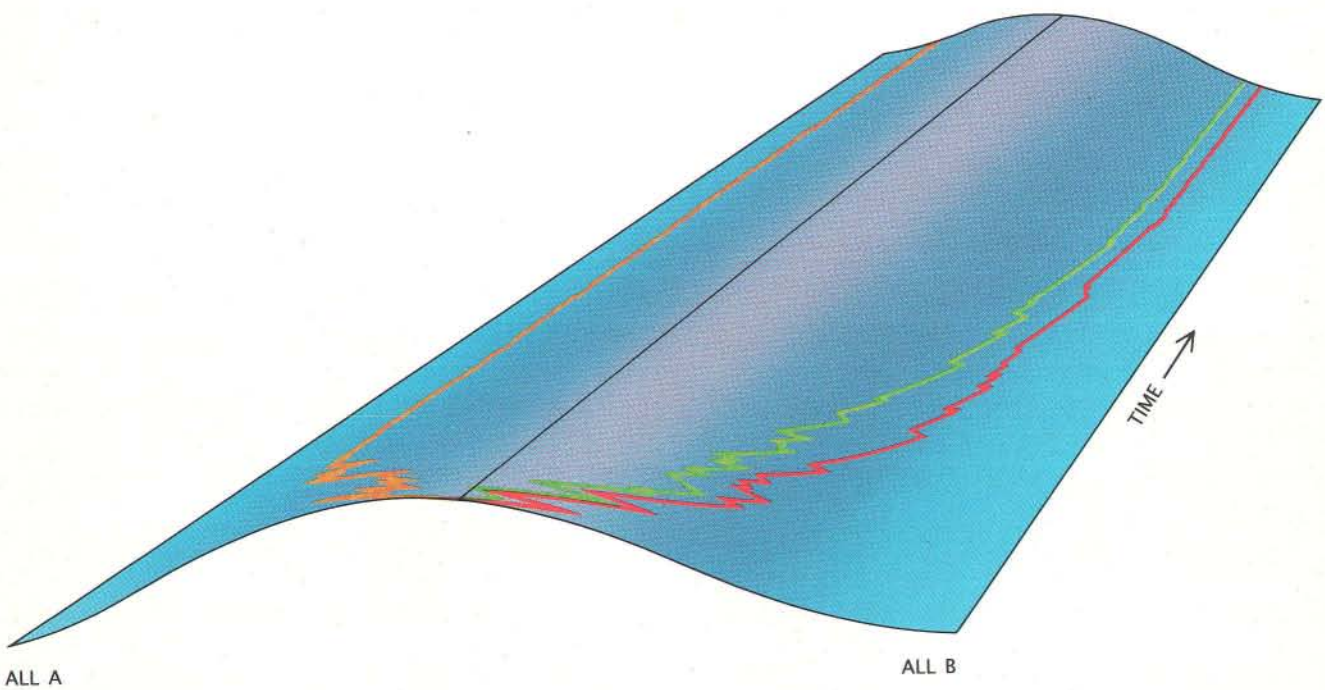
Increased production brings additional benefits: producing more units means gaining more experience in the manufacturing process and achieving greater understanding of how to produce additional units even more cheaply. Moreover, experience gained with one product or technology can make it easier to produce new products incorporating similar or related technologies. Japan, for example, leveraged an initial investment in building precision instruments into a capacity for building consumer electronics products and then the integrated circuits that went into them.

Not only do the costs of producing high-technology products fall as a company makes more of them, but the benefits of using them increase. Many items such as computers or telecommunications equipment work in networks that require compatibility; when one brand gains a significant market share, people have a strong incentive to buy more of the same product so as to be able to exchange information with those using it already.

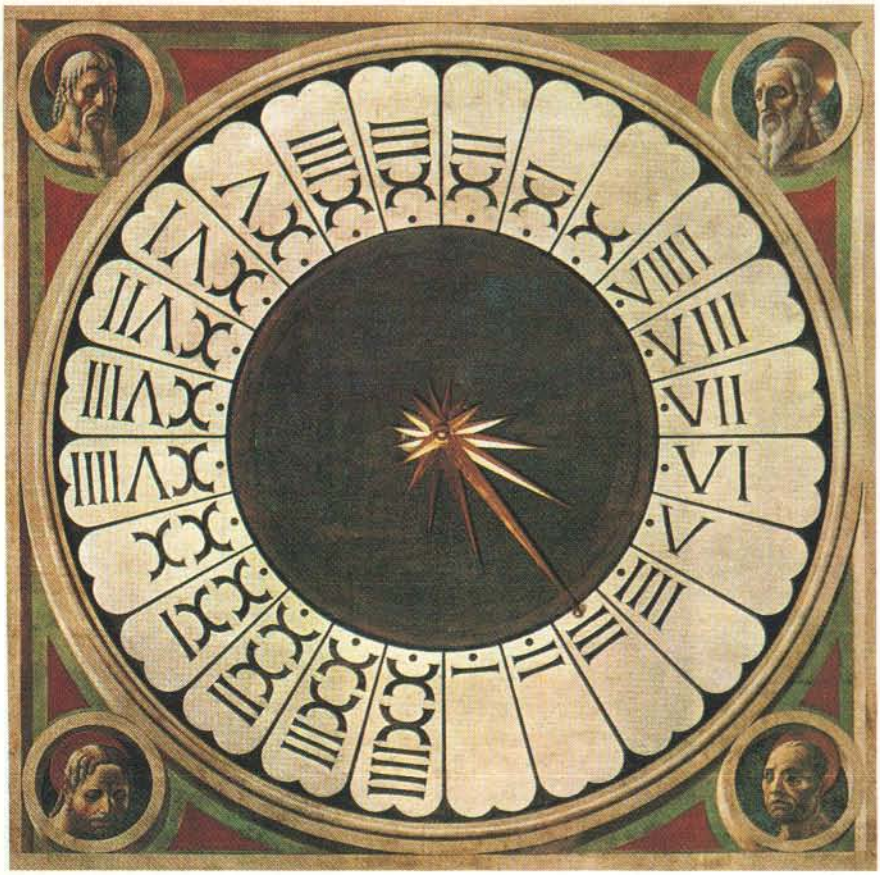
If increasing returns are important, why were they largely ignored until recently? Some would say that complicated products—high technology—for which increasing returns are so important, are themselves a recent phenomenon. This is true but is only part of the answer. After all, in the 1940's and 1950's, economists such as Gunnar K. Myrdal and Nicholas Kaldor identified positive-feedback mechanisms that did not involve technology. Orthodox economists avoided increasing returns for deeper reasons.

Some economists found the existence of more than one solution to the same problem distasteful—unscientific. "Multiple equilibria," wrote Joseph A. Schumpeter in 1954, "are not necessarily useless, but from the standpoint of *any* exact science the existence of a uniquely determined equilibrium is, of course, of the utmost importance, even if proof has to be purchased at the price of very restrictive assumptions; without any possibility of proving the existence of [a] uniquely determined equilibrium—or at all events, of a small number of possible equilibria—at however high a level of abstraction, a field of phenomena is really a chaos that is not under analytical control."

Other economists could see that



RANDOM WALK on a convex surface illustrates increasing-returns competition between two technologies. Chance determines early patterns of adoption and so influences how fast each competitor improves. As one technology gains more adherents (corresponding to motion downhill toward either edge of the surface), further adoption is increasingly likely.



FLORENCE CATHEDRAL CLOCK has hands that move "counterclockwise" around its 24-hour dial. When Paolo Uccello designed the clock in 1443, a convention for clockfaces had not emerged. Competing designs were subject to increasing returns: the more clockfaces of one kind were built, the more people became used to reading them. Hence, it was more likely that future clockfaces would be of the same kind. After 1550, "clockwise" designs displaying only 12 hours had crowded out other designs. The author argues that chance events coupled with positive feedback, rather than technological superiority, will often determine economic developments.

theories incorporating increasing returns would destroy their familiar world of unique, predictable equilibria and the notion that the market's choice was always best. Moreover, if one or a few firms came to dominate a market, the assumption that no firm is large enough to affect market prices on its own (which makes economic problems easy to analyze) would also collapse. When John R. Hicks surveyed these possibilities in 1939 he drew back in alarm. "The threatened wreckage," he wrote, "is that of the greater part of economic theory." Economists restricted themselves to diminishing returns, which presented no anomalies and could be analyzed completely.

Still others were perplexed by the question of how a market could select one among several possible solutions. In Marshall's example, the firm that is the largest at the outset has the lowest production costs and must inevitably win in the market. In that case, why would smaller firms compete at all? On the other hand, if by some chance a market started with several identical firms, their market shares would remain poised in an unstable equilibrium forever.

Studying such problems in 1979, I believed I could see a way out of many of these difficulties. In the real world, if several similar-size firms entered a market at the same time, small fortuitous events—unexpected orders, chance meetings with buyers, managerial whims—would help determine which ones achieved early sales and, over time, which firm dominated. Economic activity is quantized by individual transactions that are too small to observe, and these small "random" events can accumulate and become magnified by positive feedbacks so as to determine the eventual outcome. These facts suggested that situations dominated by increasing returns should be modeled not as static, deterministic problems but as dynamic processes based on random events and natural positive feedbacks, or nonlinearities.

With this strategy an increasing-returns market could be re-created in a theoretical model and watched as its corresponding process unfolded again and again. Sometimes one solution would emerge, sometimes (under identical conditions) another. It would be impossible to know in advance which of the many solutions would emerge in any given run. Still, it would be possible to record the particular set of random events leading to each solution and to study the probability that a particular solution would emerge under a certain set of initial conditions. The idea was simple, and it may well have occurred to economists in the past. But making it work called for nonlinear random-process theory that did not exist in their day.

Every increasing-returns problem need not be studied in isolation; many turn out to fit a general nonlinear probability schema. It can be pictured by imagining a table to which balls are added one at a time; they can be of several possible colors—white, red, green or blue. The color of the ball to be added next is unknown, but the probability of a given color depends on the current proportions of colors on the table. If an increasing proportion of balls of a given color increases the probability of adding another ball of the same color, the system can demonstrate positive feedback. The question is, Given the function that maps current proportions to probabilities, what will be the proportions of each color on the table after many balls have been added?

In 1931 the mathematician George Polya solved a very particular version of this problem in which the probability of adding a color always equaled its current proportion. Three U.S. probability theorists, Bruce M. Hill of the University of Michigan at Ann Arbor and David A. Lane and William D. Sudderth of the University of Minnesota at Minneapolis, solved a more general, nonlinear version in 1980. In 1983 two Soviet probability theorists, Yuri M. Ermoliev and Yuri M. Kaniovski, both of the Glushkov Institute of Cybernetics in Kiev, and I found the solution to a very general version. As balls continue to be added, we proved, the proportions of each color must settle down to a "fixed point" of the probability function—a set of values where the probability of adding each color is equal to the proportion of that color on the table. Increasing returns allow several such sets of fixed points.

This means that we can determine the possible patterns or solutions of an increasing-returns problem by solving the much easier challenge of finding the sets of fixed points of its probability function. With such tools economists can now define increasing-returns problems precisely, identify their possible solutions and study the process by which a solution is reached. Increasing returns are no longer "a chaos that is not under analytical control."

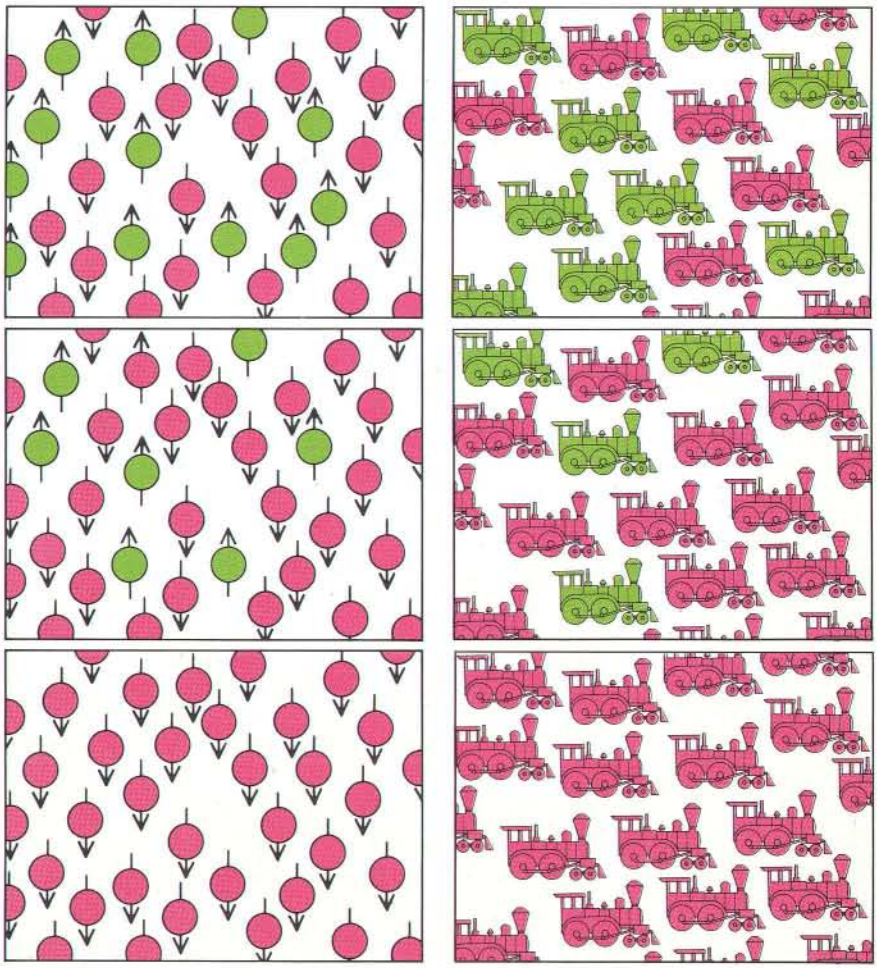
In the real world, the balls might be represented by companies and their colors by the regions where they decide to settle. Suppose that firms enter an industry one by one and choose their locations so as to maximize profit. The geographic preference of each firm (the intrinsic benefits it gains from being in a particular region) varies; chance determines the preference of the next firm to enter the industry. Also suppose, however, that firms' profits increase if they are near other firms (their suppliers or customers). The first firm to enter the industry picks a location based purely on geographic preference. The second firm decides based on preference modified by the benefits gained by locating near the first firm. The third firm is influenced by the positions of the first two firms, and so on. If some location by good fortune attracts more firms than the others in the early stages of this evolution, the probability that it will attract more firms increases. Industrial concentration becomes self-reinforcing.

The random historical sequence of firms entering the industry determines which pattern of regional settlement results, but the theory shows that not all patterns are possible. If the attractiveness exerted by the presence of other firms always rises as more firms are added, some region will always dominate and shut out all others. If the attractiveness levels off, other solutions, in which regions share the industry, become possible. Our new tools tell us which types of solutions can occur under which conditions.

Do some regions in fact amass a large proportion of an industry because of historical chance rather than geographic superiority? Santa Clara County in California (Silicon Valley) is a likely example. In the 1940's and early 1950's certain key people in the U.S. electronics industry—the Varian brothers, William Hewlett and David Packard, William Shockley—set up shop near Stanford University; the local availability of engineers, supplies and components that these early firms helped to create made Santa Clara County extremely attractive to the 900 or so firms that followed. If these early entrepreneurs had preferred other places, the densest concentration of electronics in the country might well be somewhere else.

On a grander scale, if small events in history had been different, would the location of cities themselves be different? I believe the answer is yes. To the degree that certain locations are natural harbors or junction points on rivers or lakes, the pattern of cities today reflects not chance but geography. To the degree that industry and people are attracted to places where such resources are already gathered, small, early chance concentrations may have been the seeds of today's configuration of urban centers. "Chance and necessity," to use Jacques Monod's phrase, interact. Both have played crucial roles in the development of urban centers in the U.S. and elsewhere.

Self-reinforcing mechanisms other than these regional ones work in international high-tech manufacturing and trade. Countries that gain high volume and experience in a high-technology industry can reap advantages of lower cost and higher quality that may make it possible for them to shut out other countries. For example, in the early 1970's, Japanese automobile makers began to sell significant numbers of small cars in the U.S. As Japan gained market volume without much opposition from Detroit, its engineers and production workers gained experience, its costs fell and its products improved. These factors, together with improved sales networks, allowed Japan to increase



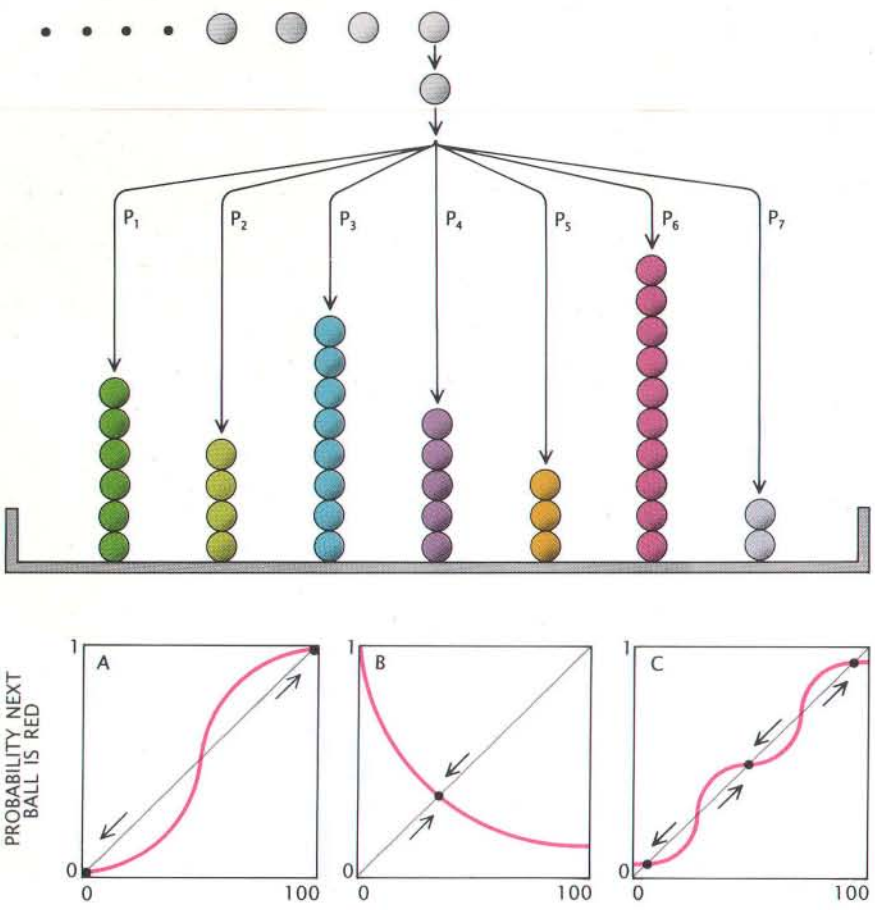
FERROMAGNETS AND REGIONAL RAIL GAUGES become ordered in much the same way. As a disordered magnetic material is cooled (*left*), the atomic dipoles inside it exert forces on one another, causing neighboring dipoles to align. Eventually all the dipoles in a sample line up, but the direction they all take (up or down) cannot be predicted beforehand. Similarly, as Douglas Puffert of Swarthmore College has shown, neighboring private railroads (*right*) in the past century adopted the same gauge to extend their range more easily. Eventually all (or most) railroads used the same gauge. Similar equations describe the behavior of these two systems.

its share of the U.S. market; as a result, workers gained still more experience, costs fell further and quality improved again. Before Detroit responded seriously, this positive-feedback loop had helped Japanese companies to make serious inroads into the U.S. market for small cars. Similar sequences of events have taken place in the markets for television sets, integrated circuits and other products.

How should countries respond to a world economy where such rules apply? Conventional recommendations for trade policy based on constant or diminishing returns tend toward low-profile approaches. They rely on the open market, discourage monopolies and leave issues such as R&D spending to companies. Their underlying assumption is that there is a fixed world price at which producers load goods onto the market, and so interference with local costs and prices by means of subsidies or tariffs is unproductive. These policies are appropriate for the diminishing-returns parts of the economy, not for the technology-based parts where increasing returns dominate.

Policies that are appropriate to success in high-tech production and international trade would encourage industries to be aggressive in seeking out product and process improvements. They would strengthen the national research base on which high-tech advantages are built. They would encourage firms in a single industry to pool their resources in joint ventures that share up-front costs, marketing networks, technical knowledge and standards. They might even foster strategic alliances, enabling companies in several countries to enter a complex industry that none could tackle alone. Increasing-returns theory also points to the importance of timing when undertaking research initiatives in new industries. There is little sense in entering a market that is already close to being locked-in or that otherwise offers little chance of success. Such policies are slowly being advocated and adopted in the U.S.

The value of other policies, such as subsidizing and protecting new industries—bioengineering, for example—to capture foreign markets, is debatable. Dubious feedback benefits have sometimes been cited to justify government-sponsored white elephants. Furthermore, as Paul R. Krugman of the Massachusetts Institute of Technology and several other economists have pointed out, if one country pursues such policies, others will retaliate by subsidizing their own high-technology industries. Nobody gains. The question of optimal industrial and trade policy based on increasing returns is currently being studied intensely. The policies countries choose will determine not only the shape of the global economy in the 1990's but also its winners and its losers.

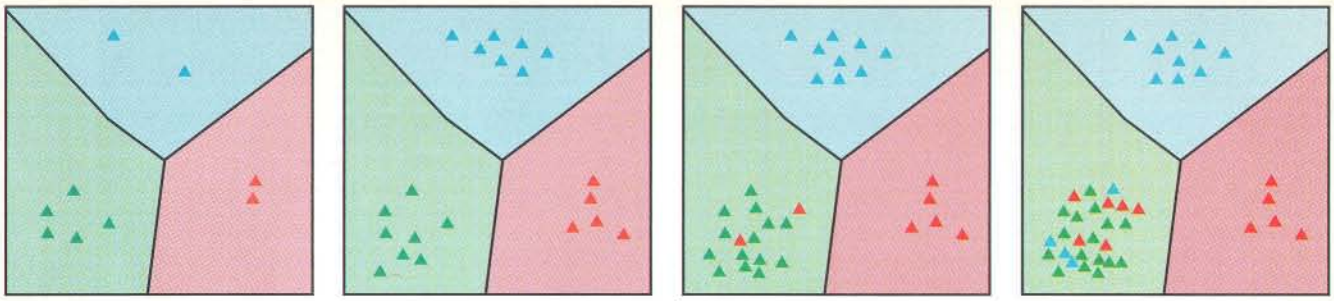


NONLINEAR PROBABILITY THEORY can predict the behavior of systems subject to increasing returns. In this model, balls of different colors are added to a table; the probability that the next ball will have a specific color depends on the current proportions of colors (*top*). Increasing returns occur in *A* (the graph shows the two-color case; arrows indicate likely directions of motion): a red ball is more likely to be added when there is already a high proportion of red balls. This case has two equilibrium points: one at which almost all balls are red; the other at which very few are red. Diminishing returns occur in *B:* a higher proportion of red balls lowers the probability of adding another. There is a single equilibrium point. A combination of increasing and diminishing returns (*C*) yields many equilibrium points.

Increasing-returns mechanisms do not merely tilt competitive balances among nations; they can also cause economies—even such successful ones as those of the U.S. and Japan—to become locked into inferior paths of development. A technology that improves slowly at first but has enormous long-term potential could easily be shut out, locking an economy into a path that is both inferior and difficult to escape.

Technologies typically improve as more people adopt them and firms gain experience that guides further development. This link is a positive-feedback loop: the more people adopt a technology, the more it improves and the more attractive it is for further adoption. When two or more technologies (like two or more products) compete, positive feedbacks make the market for them unstable. If one pulls ahead in the market, perhaps by chance, its development may accelerate enough for it to corner the market. A technology that improves more rapidly as more people adopt it stands a better chance of surviving—it has a "selectional advantage." Early superiority, however, is no guarantee of long-term fitness.

In 1956, for example, when the U.S. embarked on its nuclear-power program, a number of designs were proposed: reactors cooled by gas, light water, heavy water, even liquid sodi-



COMPANIES CHOOSE LOCATIONS to maximize profits, which are determined by intrinsic geographic preference (shown by color) and by the presence of other companies. In this computer-generated example, most of the first few companies settle in the green region, and so all new companies eventually settle there. Such clustering might appear to imply that the green region is somehow superior. In other runs of the program, however, the red and blue regions dominate instead.

um. Robin Cowan of New York University has shown that a series of trivial circumstances locked virtually the entire U.S. nuclear industry into light water. Light-water reactors were originally adapted from highly compact units designed to propel nuclear submarines. The role of the U.S. Navy in early reactor-construction contracts, efforts by the National Security Council to get a reactor—any reactor—working on land in the wake of the 1957 *Sputnik* launch as well as the predilections of some key officials all acted to favor the early development of light-water reactors. Construction experience led to improved light-water designs and, by the mid-1960's, fixed the industry's path. Whether other designs would, in fact, have been superior in the long run is open to question, but much of the engineering literature suggests that high-temperature, gas-cooled reactors would have been better.

Technological conventions or standards, as well as particular technologies, tend to become locked-in by positive feedback, as my colleague Paul A. David of Stanford has documented in several historical instances. Although a standard itself may not improve with time, widespread adoption makes it advantageous for newcomers to a field—who must exchange information or products with those already working there—to fall in with the standard, be it the English language, a high-definition television system, a screw thread or a typewriter keyboard. Standards that are established early (such as the 1950's-vintage computer language FORTRAN) can be hard for later ones to dislodge, no matter how superior would-be successors may be.

Until recently conventional economics texts have tended to portray the economy as something akin to a large Newtonian system, with a unique equilibrium solution preordained by patterns of mineral resources, geography, population, consumer tastes and technological possibilities. In this view, perturbations or temporary shifts—such as the oil shock of 1973 or the stock-market crash of 1987—are quickly negated by the opposing forces they elicit. Given future technological possibilities, one should in theory be able to forecast accurately the path of the economy as a smoothly shifting solution to the analytical equations governing prices and quantities of goods. History, in this view, is not terribly important; it merely delivers the economy to its inevitable equilibrium.

Positive-feedback economics, on the other hand, finds its parallels in modern nonlinear physics. Ferromagnetic materials, spin glasses, solid-state lasers and other physical systems that consist of mutually reinforcing elements show the same properties as the economic examples I have given. They "phase lock" into one of many possible configurations; small perturbations at critical times influence which outcome is selected, and the chosen outcome may have higher energy (that is, be less favorable) than other possible end states.

This kind of economics also finds parallels in the evolutionary theory of punctuated equilibrium. Small events (the mutations of history) are often averaged away, but once in a while they become all-important in tilting parts of the economy into new structures and patterns that are then preserved and built on in a fresh layer of development.

In this new view, initially identical economies with significant increasing-returns sectors do not necessarily select the same paths. Instead they eventually diverge. To the extent that small events determining the overall path always remain beneath the resolution of the economist's lens, accurate forecasting of an economy's future may be theoretically, not just practically, impossible. Steering an economy with positive feedbacks into the best of its many possible equilibrium states requires good fortune and good timing—a feel for the moments when beneficial change from one pattern to another is most possible. Theory can help identify these states and times, and it can guide policymakers in applying the right amount of effort (not too little but not too much) to dislodge locked-in structures.

The English philosopher of science Jacob Bronowski once remarked that economics has long suffered from a fatally simple structure imposed on it in the 18th century. I find it exciting that this is now changing. With the acceptance of positive feedbacks, economists' theories are beginning to portray the economy not as simple but as complex, not as deterministic, predictable and mechanistic but as process-dependent, organic and always evolving.

FURTHER READING

MARKET STRUCTURE AND FOREIGN TRADE. Elhanan Helpman and Paul Krugman. The MIT Press, 1985.

PATH-DEPENDENT PROCESSES AND THE EMERGENCE OF MACRO-STRUCTURE. W. Brian Arthur, Yu M. Ermoliev and Yu M. Kaniovski in *European Journal of Operational Research*, Vol. 30, pages 294–303; 1987.

SELF-REINFORCING MECHANISMS IN ECONOMICS. W. Brian Arthur in *The Economy as an Evolving Complex System*. Edited by Philip W. Anderson, Kenneth J. Arrow and David Pines. Addison-Wesley Publishing Co., 1988.

PATH-DEPENDENCE: PUTTING THE PAST INTO THE FUTURE OF ECONOMICS. Paul David. I.M.S.S.S. Tech Report No. 533, Stanford University; November, 1988.

COMPETING TECHNOLOGIES, INCREASING RETURNS, AND LOCK-IN BY HISTORICAL EVENTS. W. Brian Arthur in *The Economic Journal*, Vol. 99, No. 394, pages 116–131; March, 1989.

THE AMATEUR SCIENTIST

When a polymer sheet is stretched, it may "neck" long before it snaps

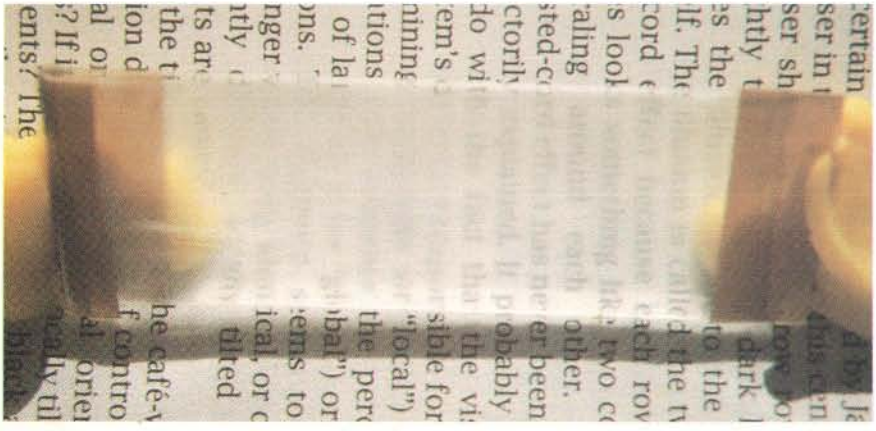
by Jearl Walker

Many polymer sheets, including some common plastic food wraps, behave peculiarly when they are stretched. They neither snap like a thread nor expand like a rubber band. Instead they first strongly resist the stretching and then suddenly yield by narrowing in thickness or in width (perpendicular to the direction of stretch), or in both dimensions. The narrowing is called necking or cold drawing. (The second term refers to a similar narrowing that a hot metal rod undergoes when it is drawn.)

Test a length of stretchable plastic sheet by pulling on the two ends. At first you need to pull hard to get any movement at all, but then the plastic abruptly stretches and necks. Once it has necked, the plastic becomes easier to stretch, and the narrowing begins to travel toward the ends. After the necking has spread appreciably, though, you must again pull hard, and finally the plastic rips somewhere in the narrowed region.

Before I explain the mechanics of necking, let me mention a rather puzzling effect of the phenomenon. Some polymer sheets are transparent but murky. To read a printed page through one, you have to hold the sheet close to the page; as you move the sheet farther from the page, the print soon becomes too obscure to read. The murkiness is caused primarily by the scattering of light by the molecules of the sheet. The greatest distance at which you can read through a sheet is one measure of how strongly the light is scattered.

Now, if you stretch the sheet and it necks in thickness, the light passes through less material—fewer molecules—and the net scatter should be less; you should be able to hold the sheet farther from the page and still be able to read the print. As logical as the argument seems, clearly it does not hold true in the situation pictured below. To make the photograph, I stretched and necked half of a polymer strip, laid the strip flat on a microscope slide and then, with modeling clay, propped up the slide over a printed page at an angle, so that the distance between the page and the strip varied. In the section of the unnecked region farthest from the page the print is blurred but readable, whereas in the necked region the print blurs out completely even where the strip is much closer to the page. The reason is subtle, as will appear below.



The necked region of a polymer sheet obscures print

A polymer is a large molecule built up by the repetition of some basic chemical unit called a monomer. The many different polymers encountered in daily life present a wide range of chemical structures and consequently of mechanical and optical properties. From the subset that displays necking, I chose two examples to study. One is polyethylene, most commonly found in the kitchen in the form of storage bags or sheets—often self-sticking—for wrapping foods. The monomer of polyethylene is a simple array of two carbon atoms and four hydrogens. In a sheet of polyethylene, regions where the long polymer molecules form tiny crystals are separated by amorphous regions that lack any organization. The other polymer sheet I chose to study is Parafilm, which is ubiquitous in biological and chemical laboratories, where it serves to seal off beakers and other containers. Parafilm is a mixture of wax and polyethylene.

Necking is in large part a consequence of the orientation of polymer molecules. Consider a polymer sheet that is necking in thickness, and imagine that you can see the molecules in the sheet. Initially they might be organized on a small scale—partially crystallized or somehow aligned by the manufacturing process—or they may have completely random orientations. As you pull, the sheet is able to stretch only if the molecules can be made to turn or shift to line up with the direction of the pull and thereby accommodate the increase in length. At first their chemical bonds resist the reorientation, and the sheet stretches only slightly. But once the pull reaches some critical value, the molecules in the weakest region of the sheet surrender, break their weaker bonds, slip over one another and move into alignment [*see top left illustration on opposite page*]. The sheet grudgingly yields by sacrificing thickness for length in that region. The sheet does gain strength, though, in the sense that the reoriented molecules have shorn their weakest connections.

If you continue to pull, you align more of the molecules in the necked region and further diminish the thickness. Once many of the molecules are

"The Amateur Scientist" and A. K. Dewdney's newly titled "Mathematical Recreations" will appear in this space in alternate months.

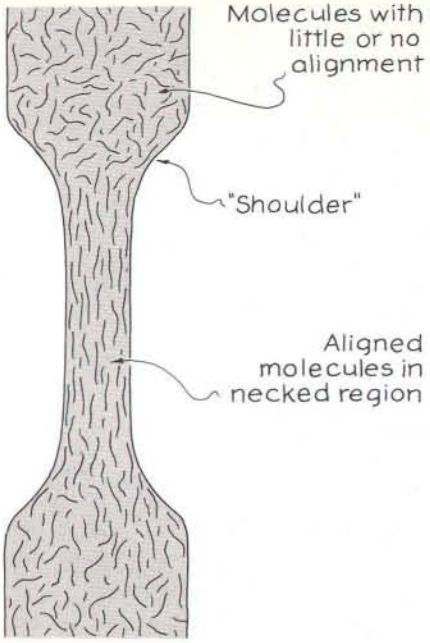
aligned, the thickness is at its narrowest, and then the bonds among the molecules are too strong to allow any further yield. If you keep pulling, the molecules in the "shoulders" of the necked region are the next to submit, which spreads the necking along the sheet in the direction of your pull. When the full sheet is necked, the "hardened" nature of the molecular interconnections throughout the sheet requires that you once again pull strongly on the sheet to stretch it still more, and soon the sheet snaps rather than give in much further.

To prepare Parafilm to be necked, I cut a rectangular strip with scissors, laid the strip out flat on a table and then tacked each end down with sturdy packaging tape. Next, with ruler and pen, I marked off lines that ran across the width of the strip and were spaced two millimeters apart. Then I transferred the strip to a laboratory jack, pressing each taped end firmly to a face of the jack. (You could substitute a household vise.) The strip extended vertically between the jack faces; it was straight but was under no tension.

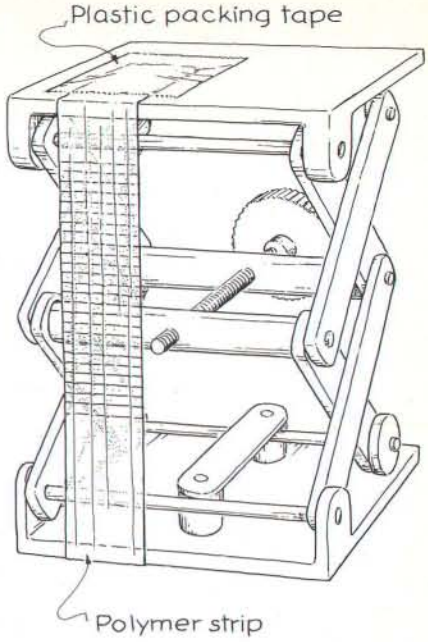
I recorded the vertical length and the left-to-right width of the strip. To measure its thickness with a micrometer, I first backed off the instrument's mobile prong enough to allow the strip to slip through the prongs and then gradually tightened the prongs while also gently moving the micrometer around. When the prongs were close enough to catch on the strip, I backed them off just enough to eliminate the catch and recorded the micrometer setting. I repeated the measurement several times and averaged the results. The strip was approximately .14 millimeter thick.

Now I began to turn the jack screw, moving the two faces of the jack apart and thereby stretching the strip. The turning was difficult at first, but when I had stretched the strip by about 9 percent, it suddenly yielded: it necked in thickness within a narrow band that ran across its width. I found that inked lines in the necked region were separated by an additional 25 percent; the lines elsewhere showed no extra separation.

As I continued to turn the jack screw, the necking inched toward the two faces of the jack, stretching apart lines along the way. The progress was not uniform: after each turn of the screw, the shoulders of the necked region were marked with "islands" of unstretched material that were visibly different from the stretched material surrounding them. With each turn of the screw I measured the length, width and thickness of the strip and the separation of lines at various spots along the strip. When I had stretched the strip by about 90 percent, all the lines on the strip displayed additional separation.



Molecular alignment in necking



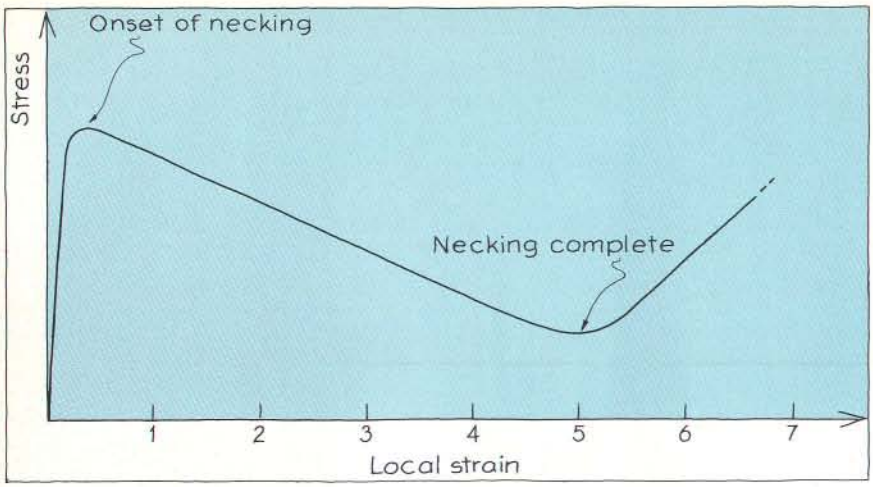
How to stretch a polymer sheet

When the stretch reached 150 percent (that is, when the strip had been stretched to two and a half times its original length), the lines in the initially necked region were as much as nine millimeters apart (which corresponds to a stretch of 350 percent) and the thickness was only .064 millimeter. Elsewhere the lines were 3.5 millimeters apart (the material between them had been stretched by 75 percent) and the thickness had been somewhat reduced. The stretching and thinning were obviously spreading along the strip. With one more turn of the screw, the strip began to tear in the necked region, probably because of a nick I had left along the side of the strip with the scissors. (I could have sealed the tear with a small patch of packaging tape but had not done so.)

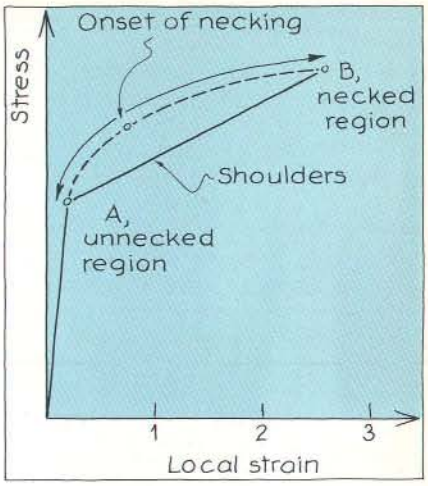
The stretching of the strip can be followed with a graph in which "conventional stress" is plotted against "strain" [*see top left illustration on next page*]. The strain, which is the same as the extent of stretch, is the ratio of the change in the length of the strip to its initial, unstretched length. For example, when the length of the strip is doubled, the stretch is 100 percent and the strain is 1.0. The conventional stress is the ratio of the pull on the strip to the area of a cross section of the strip along the edge of a face of the jack [*see bottom illustration on next page*]. That area did not change in the course of the experiment, and so any change in the conventional stress reflected the change in the strength of the pull.

Because I could not measure the pull, I have graphed only my subjective observations. As I first turned the screw, the stress climbed and the strip stretched marginally. When the strain reached .09, the strip suddenly necked and lengthened, thereby relieving the pull by the jack and decreasing the stress on the strip. As I turned the screw more, increasing the strain, more of the strip necked and the stress decreased even further. If I had been able to neck the full strip without its tearing because of an accidental nick (I came close with several samples), the stress would have begun to increase again as strongly linked molecules resisted further stretching. Eventually the stress would have snapped the strip like a thread.

Before necking appears, the conventional stress applies to the entire strip and also to any given section. After necking, however, the stress in the necked region is larger than the conventional stress because the cross-sectional area is reduced there. Such local stress is called the true stress; it is the ratio of the pull on a region to the region's cross-sectional area. The strain in the region is said to be the local strain.



The curve for conventional stress



The curve for true stress

In the illustration at the right above, true stress is plotted against local strain before and after necking appears. Prior to necking, every section of the strip undergoes the same stress and strain, and the point representing those values climbs up into the broken part of the curve. Just as the strip necks, point *A*, representing unnecked regions, slides back down the curve and point *B*, representing the necked region, climbs higher on the curve: the necked region is under more stress and strain than the unnecked regions. The line directly connecting *A* and *B* indicates stress and strain through the shoulders.

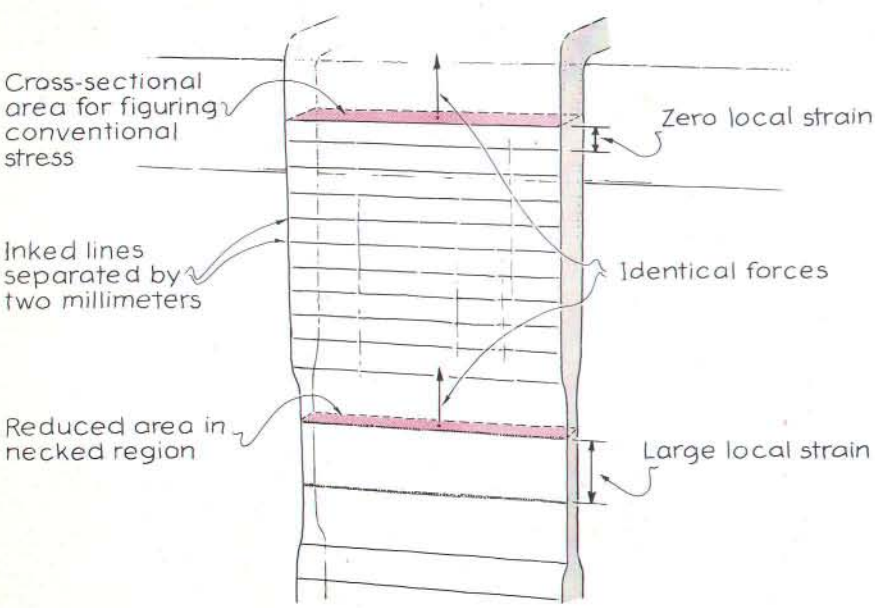
I repeated the stretching experiments with strips of Glad Cling Wrap, a clear polyethylene sheet. Like Parafilm, a strip of the food wrap strongly resisted the initial stretching and then suddenly necked. Unlike Parafilm, however, it necked in its left-to-right width rather than in thickness. The narrowing began in one place and then gradually spread up and down the strip, giving it a distorted hourglass shape. Just before the strip ripped, it was stretched by 180 percent and the initial width of three centimeters had been reduced to .9 centimeter at the narrowest region. Inked lines there indicated a local strain of about 4.0.

I now return to the matter of reading print through a polymer sheet. The matter was discussed in 1973 by David Miller of the Beth Israel Hospital and Harvard University and George B. Benedek of the Massachusetts Institute of Technology in their book *Intraocular Light Scattering*. They called the effect "the nude in shower phenomenon." (I am grateful to Craig F. Bohren of Pennsylvania State University for pointing out the reference.)

To see how the nickname for the effect got its start, imagine that you watch a bather through a rigid shower enclosure or curtain made of textured plastic. If the person is close to the plastic, body features are readily visible, but if the person is farther from the plastic, the features are too murky to distinguish. A similar dependence of visibility on distance can be observed with a strip of common transparent plastic tape. Hold the tape (adhesive side up to prevent sticking!) just above this page, and then gradually move it upward. As its distance from the page increases, the words begin to blur and become unreadable.

To understand the blurring, consider a dot on the page. When light travels from the dot up through the tape, it scatters from the tape's molecules [*see drawing at left in illustration on opposite page*]. The scattering spreads each original ray into a bundle of rays forming a narrow cone centered on the direction of the original ray. The cone can be characterized by its half-angle: the angle between the most severely scattered rays in the cone and the direction of the original ray.

Suppose that the tape is just above the dot and that you look down on it from a distance of at least 40 centimeters with one eye closed. Your open eye then intercepts rays that are approximately perpendicular to the tape and that have been scattered by molecules along, or adjacent to, the direct line to the dot. Rays that scatter from molecules farther out miss your eye and do not contribute to your vision. To perceive the source of the inter-



Factors that determine stress and strain

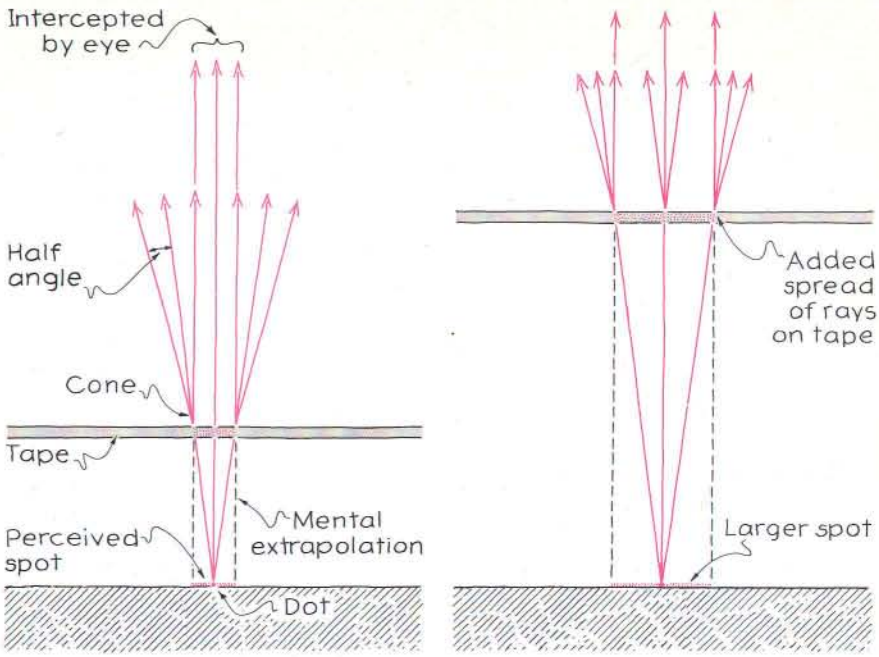
cepted rays, your brain automatically extrapolates them back to the paper, where they appear to have originated from a small spot centered on the actual dot. The spot will not be as crisp as the dot, but it is still recognizable. Its radius is approximately equal to the product of the dot-to-tape distance and the tangent of the half-angle of the scatter cones.

Now lift the tape a few millimeters. As is shown in the drawing at the right in the illustration, the half-angle of each scatter cone is unchanged, but the extra dot-to-tape distance spreads the light on the tape, and so you intercept rays from a somewhat larger region of the tape around the direct line between the dot and your eye. As you extrapolate the rays backward, they appear to originate from a somewhat larger spot on the page. The spot is now less distinct; it is harder to recognize it as being a dot.

Suppose there is a second dot next to the first one. When the tape is near the dots, the perceived spots are small and sharp enough to be distinguished as being separate. As you lift the tape and the spots widen, they eventually overlap too much to be distinguished. So it goes with the nude behind the textured plastic: when the figure is close to the plastic, details are distinct and recognizable, but when the figure is farther away, the details blur out.

While experimenting with Parafilm, I noticed that it obscured the details of a printed page seen through it, just as tape does. I assumed that the extent of blurring must depend on the thickness of the sheet. For a certain thickness, each ray from a given detail must pass through a certain number of molecules and thus be spread into a cone with a certain half-angle, and that half-angle should determine how far from the page I can hold the sheet and still make out the detail. Presumably, if I thinned the sheet by necking it, the light would be scattered by fewer molecules and be spread into a narrower cone, and I should then be able to distinguish a detail with the sheet farther away from it.

To test the idea, I placed a small section of unstretched Parafilm over the hole in a mechanical-drawing template and positioned the hole over three small dots I had penciled on a sheet of paper. The dots formed a corner and were separated vertically and horizontally by a millimeter. I lifted the template while looking at the dots through the Parafilm. When the Parafilm was about two centimeters from the paper, the dots blurred together. I repeated the observation a number of times and averaged the "blur-out" heights.



How plastic tape spreads light rays

I next substituted a small section of Parafilm that had been necked with the jack. To my surprise, I found that the dots blurred out when the section was only half a centimeter above them. Moreover, the two dots aligned with the section's direction of stretch blurred out about a millimeter lower than the two dots oriented perpendicular to that direction. (Of course, that discrepancy could have been caused by an unequal separation between the dots in the horizontal and vertical directions, but the discrepancy persisted even when I turned the paper around my line of sight by 90 degrees.) Although my observations were certainly crude, they indicate that when the Parafilm necks, it scatters light in cones that are roughly four times broader than the cones created when the sheet is unstretched.

I think the increase in scattering results from the alignment of molecules that is brought about by the necking. When light travels through a transparent material in which the molecules are randomly oriented or in which their organization is on a scale much smaller than the wavelength of light, each molecule is said to scatter the light independently of the other molecules. In such a case, any light that is scattered out of the forward direction is likely to be canceled by light scattered in the same direction by another molecule. (That is, each crest of one wave falls on a valley of the other wave; the waves interfere destructively.) The half-angle of a scatter cone is thereby kept small.

If, instead, there is some ordered arrangement of the molecules, and if the spacing associated with the arrangement is about the size of the wavelength of light, the molecules scatter not independently but in an organized way. When two light waves are then scattered in the same direction, the crests of one wave may not fall precisely on the valleys of the second wave; the cancellation of light scattered to the side is less complete, and the half-angle of each scatter cone is broader. Apparently, when I aligned the molecules by necking the Parafilm, I organized the scattering, widened the scatter cones and increased the blurring by the sheet.

FURTHER READING

X-RAY DIFFRACTION STUDIES OF THE STRETCHING AND RELAXING OF POLYETHYLENE. Alexander Brown in *Journal of Applied Physics*, Vol. 20, No. 6, pages 552–558; June, 1949.

THE NECKING AND COLD-DRAWING OF RIGID PLASTICS. P. I. Vincent in *Polymer*, Vol. 1, No. 1, pages 7–19; 1960.

ON THE EXTENSION OF THE NECK OF POLYMER SPECIMENS UNDER TENSION. G. I. Barenblatt in *Journal of Applied Mathematics and Mechanics*, Vol. 28, No. 6, pages 1264–1276; 1964.

INTRAOCULAR LIGHT SCATTERING: THEORY AND CLINICAL APPLICATION. David Miller and George Benedek. Charles C Thomas, Publisher, 1973.

BOOKS

Good breeding, the raven's yell, the spillproof fuel, icons of science

by Philip Morrison

FOOD CROPS FOR THE FUTURE: THE DEVELOPMENT OF PLANT RESOURCES, by Colin Tudge. Basil Blackwell, 1988 ($49.95; paperbound, $19.95).

The topic is urgent: "the conversion of edible plants into food crops capable of feeding 10 billion people." The author is an English science writer of experience and learning. His gift for clear exposition and his width of biological vision make his brief book an inviting guide to the up-to-date applied science of plant breeding. Without even a single diagram, he brings the reader to a level of understanding not to be gained from most longer and bristlier texts. One reason is his sharpness of focus: we learn easily when purpose and context are never far from view.

Plants and animals adopt different life strategies: plants stay in the same place; animals move to get where they want. Locomotion implies swift coordination and good control of form; plants need resilience, and so for them form is far more flexible. A mighty oak tree in parkland and a twisted shrub lodged in some rocky crevice could have come from the same acorn. We all know that both sexual reproduction (via seed) and asexual reproduction (from a variety of parts—tubers, suckers, bulbs and even stem cuttings) come easily to many plants. Both methods have been the means of breeding our plant symbionts.

The artisan breeders, men and women innocent of Mendel and Morgan, made enormous change by long selection, both witting and incidental. Grain that was harvested, for example, was grain that had not fallen to the ground, and so over time the varieties in the field became those that held their seed well—maize above all. With more art, the breeder favored what was wanted, purged the crop of unwanted traits and left a legacy of wonders.

There are distinct approaches to plant breeding, determined by the plant itself. Some species are natural inbreeders; wheat and barley "breed true," quite able to pollinate themselves. Crosses for new varieties given some hopeful traits, although taxing logistically (you must select among a couple of million individuals over a decade), are conceptually simple. The selected plants become more and more uniform.

Millet is one of the outbreeders and is consequently highly variable, concealing very well the outward signs of its inner genetic nature. Breeding such crops is an "exercise in compromise." In each of many test plots the plants sown are barred from pollination except with their similarly selected plot fellows. The breeder will select for many traits, combing some of the test plots for short stalks and some for tall, some for big heads, some for big seeds. After a dozen years and more, a large number of distinct varieties will have formed. Within each plot the plants grow more and more alike, while the different plots become more distinct. Here modern work packs into a decade or so the field-by-field actions of alert farmers over millennia. Finally, breeding may well forgo sexual reproduction and transmit the selected plants by asexual routes, to yield a clone of near-identical copies of the wanted model.

Sometimes a plant reproductive cell survives an abnormal cell division, in which the chromosomes have not been neatly split. The chromosome set passes down in multiple; the plant is polyploid. The pattern is not uncommon in nature; such plants are infertile save with a matching counterpart. The process amounts to a gentle form of speciation. (Most sugar cane grown today is to be traced to infertile but vigorous crosses that were made 40 years ago between cane varieties with differing chromosome sets. The chromosome set of these hybrids is bizarre, variable and ineffective, but that does not matter to the grower, for these plants are reproduced asexually as clones from bits of the cane.) The process can be induced deliberately, as can the changes in isolated genes called mutation.

Every higher organism is a family of cells, all of which carry the same ancestral genetic message. In animals, genes are suitably switched off to orchestrate development, so that the few but totipotent embryo cells can differentiate into a hundred different cell types specialized to form such tissues as skin, marrow and muscle; once they switch off, genes cannot be turned on again. In plant cells the opened developmental switches can be closed again simply by placing the cells in the right hormonal environment. This is perhaps "as profound as any other difference between the two kingdoms."

No sheep has yet been regenerated from a single hoof, whatever might be done with starfish arms. But it is all of 40 years since F. C. Steward first grew a complete plant, a carrot, from a single cultured root-hair cell; orchids were cloned from single cells too, at about the same time. By now orchid cloning is an industry that each year ships four million fine virus-free plants. Breeding to improve important slow-growing tree crops, such as oil palm and coconut, is dauntingly tedious; a dozen generations of crossings, 20 years at a stage, would try our patience. But good tree varieties are being multiplied wholesale in the "ultimate monoculture": whole groves planted with a few prized clones.

The next step is as evident as it is fantastic. For wheat or corn or potatoes, the product humans consume is a considerable fraction of the plant biomass. That is not at all true for the plants that are our chemists—those that make dyes and flavors and drugs. Only a small proportion of the cells of any such plant produce what we want. Why then grow the whole plant? Cell culture is the alternative: it "offers control—of supply, of quality, and . . . of legalities." Productive and "leaky" cell types must first be found. Cells once cultured are not sacrificed in harvesting, like maize or chickens, but are husbanded like milch cows. They are steadily milked of their product, which may well be self-toxic, by chemical separation, while they thrive within a nutrient flow, guarded against infections by antibiotics. One rolled-up green mat of leaf cells thriving perpetually in its sealed tank might rival a modest farm.

In Tokyo right now they culture the root cells of a plant that yields a prized traditional red dyestuff, sale-

able at $4 a gram. Biotechnologists estimate that any plant product able to bring as much as a dollar or two per gram is a good candidate for economic cell culture.

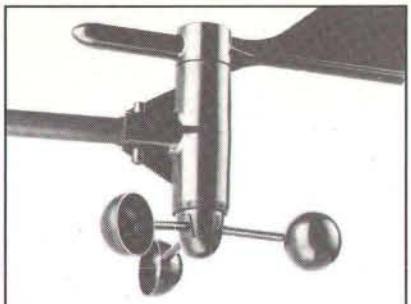
It is the virtue of this book that it treats genetic engineering in a rich context. With molecular mastery one need not use even one cell of the desired plant; instead only a few of the plant's genes need be supplied, recombined by the molecular art to inform the DNA of some easily cultured bacterium or yeast. One specific hope is to string into the chromosomes of wheat the enzymatic recipes to give that vital grain what clover has: a way of fixing nitrogen from thin air.

What will come of all this? Soon the short-stalked wheats and rices of the Green Revolution that are not beaten down by the rains may go over to strains as tall as the wheats of Breughel's day but bearing long, grain-heavy "heads... like pampas grass, and... stalks... like pokers." Potato breeders, drawing on the old Inca varieties for flavor and color, might arrange skin color to mark qualities: you would choose "a blue... strong-flavoured and floury... or an orange strong and waxy.... This would not merely be caprice. It would be good for the world's... nutrition, if people were to eat more potatoes... made as interesting as possible."

An opening chapter or two survey freshly the present barnful of crops, familiar or exotic, that provide the baseline for improvement, which will surely include the spread of some less exploited plants such as the winged bean and the quinoa. But the entire task of crop improvement is societal; it far transcends genetics. Excellent eminently usable plants are not yet crops. The author, never polemic, closes with a reminder of the necessity for justice within all these profound changes, of the task of funding research to be done on behalf of farmers too poor to pay for it, of the imperative to use science well. One precondition is a measure of public understanding; books like this point the way. The closing words are hopeful and plain: "The world could be fed."

RAVENS IN WINTER: A ZOOLOGICAL DETECTIVE STORY, by Bernd Heinrich. Summit Books, 1989 ($19.95).

It was this Vermont zoologist who told us so compellingly a few years ago of the economy of the bumblebees. In the sunny summer meadows of the northern woodlands, those furry small capitalists are programmed in the nectar market to maximize return,



SERIOUS ABOUT THE WEATHER? NOW YOU CAN AFFORD A PERSONAL WEATHER STATION!

The new WeatherPro weather station gives you the local weather at your fingertips. Easy to install and simple to operate, the WeatherPro includes a weather computer, remote precision wind direction vane, wind speed sensor, external temperature probe, mounting hardware and 40' of cable-**all for only $179!**

- WIND SPEED
- TEMPERATURE
- WIND DIRECTION
- TEMPERATURE HI/LO
- RAINFALL (OPTIONAL)*
- WIND GUST
- WIND CHILL
- TIME
- AUTO SCAN
- 1 YEAR WARRANTY
- 14-DAY MONEY-BACK GUARANTEE

DIGITAR WEATHERPRO
WEATHER STATION: ONLY $179!
ORDER TODAY:1-800-678-3669, SC
M-F 7AM-5:30 PM Pacific Time

*Automatic-emptying electronic rain gauge-$49.95
Add $5.00 for shipping. CA residents add sales tax.
Fax 1-415-732-9188 • M/C and VISA

DIGITAR®
3465 DIABLO AVE, HAYWARD, CA 94545

Authors...
LOOKING FOR A PUBLISHER?
Learn how to have your book published.

You are invited to send for a free illustrated guidebook which explains how your book can be published, promoted and marketed. Whether your subject is fiction, non-fiction or poetry, scientific, scholarly, specialized, (even controversial) this handsome 40-page brochure will show you how to arrange for prompt publication.

Unpublished authors, especially, will find this booklet valuable and informative. **For your free copy, write to: VANTAGE PRESS, Inc. Dept. F-53 516 W. 34 St., New York, N.Y. 10001**

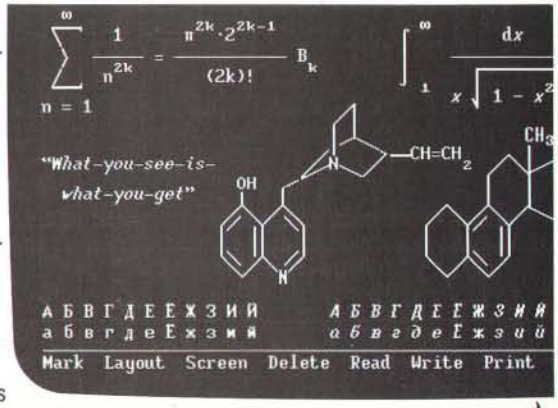
NOW!
full-featured scientific word processing at one-third the cost!

You'd have to spend more than three times the cost of **ChiWriter** to get a program nearly as powerful.

ChiWriter gives you all the features of an advanced scientific word processor—and more! The "what-you-see-is-what-you-get" screen display lets you enter and edit text and scientific notation exactly as you want it on your printout. Add easy math editing, font design, auto box mode, macros, and foreign language capability—and you have a genuine bargain at only $149.95.*

Call today to order your **ChiWriter** package. Full 30-day money back guarantee. Bankcards welcome.

***Only $149.95 for basic program plus shipping and handling.**
$74.95 for laser support. 20% educational discount.
To order, call toll free
1-800-736-8886



ChiWriter
horstmann software

FEATURES
- Screen display corresponds exactly to printout—no encoding or preview mode.
- Easy formula editing is fully integrated into the program
- Symbols and formulas can be anywhere in text
- All standard math and scientific symbols included
- Font design program—modify our symbols or design your own
- Spell checker included
- For IBM® PC or compatible systems

HORSTMANN SOFTWARE DESIGN CORPORATION
4 North 2nd. St., Ste. 500 P.O. Box 1807 San Jose, CA 95109-1807, USA Phone (408) 298-0828 Fax (408) 298-6157

IBM is a registered trademark of International Business Machines Corporation.

foraging most frugally in the short season to eke out an essential energy profit. Now he is back from watching the same Maine woods in snowy winter, with a extraordinary tale about the raven, the big, black, imposing corvid (not the common crow), found circumpolar from the Arctic to the mountains of Central America.

Paragons of keen-eyed flight, the "brains of the bird world," loquacious and varied of call even to the mimicry of speech, ravens are actively at home everywhere they range. Above all they follow the big killers—the polar bears, wolf packs, coyotes, Viking warriors and executioners. They were the garbage crew of old London; there they gorged, we are told, on the unburied corpses left by the Great Fire. Today ravens are making a return to New England as the woods fill in and roadkills and garbage dumps multiply. They live on any meat that they can tear, and so they do hunt insects and lizards and mousy creatures and smaller birds, but they specialize in the big bonanzas of carrion opened and left haphazardly behind by more powerful predators. (A raven Heinrich reared managed to eat a whole dead gray squirrel even though the bird could not penetrate the skin. It removed the flesh through the animal's mouth; the skin was discarded quite clean, with the fur on the *inside*.)

This time Heinrich tells a more personal story, an evening-after-evening field journal of puzzlement and unstinted endeavor, with clarifying chapters of later analysis. It is a paradigm of process, the search for scientific evidence. He had to enter a new discipline, turned from bumblebee economist to raven sociologist. If Economic Bee gathers daily, each buzzy optimizer for herself and all for the nest under the invisible hand, the much subtler ravens, known as pair-by-pair solitary nesters, are exceptional. Against all the odds they are altruists, cooperators—not merely with their offspring, but with crowds of raven strangers. Why?

These years of arduous research began in 1984 with a chance insight. Heinrich grew up in western Maine; now he owns a tarpaper shack there at the edge of that bee-loud clearing in the forest, a steep and weary half-mile from the road. Ravens were visitors now and then, by one's and two's. One October day he heard a raven call, and then many, from across the ridge. When he appeared there, 15 ravens and more flew off; they had been at a half-hidden moose carcass left by a poacher, some of the meat still fresh. He took a cut and made a small fire up by the trees. "I wait to watch the birds. There is no greater pleasure than eating roasted moose while resting under a spruce and contemplating ravens." Contemplation brought a question: How could so many ravens have found the meat so soon?

The answer sounded wildly in the air; raven finders recruit guests with a special loud and high-pitched call that is unlike any made away from a carcass. "I call it simply a 'yell'.... I was awed." Only active announcement could explain the banquet in that place and season. Vultures flock to a kill because they see the steep descent of the finder; that cannot be hidden. But why should a raven—who might dine alone for weeks on the rare find—share it with strangers? Even if it is good for ravenkind to share chance finds, how could such a pattern evolve through the selection of individuals?

Heinrich's dream of a fine letter for *Science* in two weeks' work was a hundredfold underestimate. The first test was to record some yelling and then to play it back, food absent. Heinrich lugged bait and electronics up the trail. Two birds came by during the next days of watching; they dined again and again, in silence. Hypothesis: A stingy bait is not worth raven publication. A big dead goat is a heavy load up the trail, but it was borne. No bonanza; maybe they don't like goat. Seek out moose meat.

A year goes by with conflicting results, leaving nine hypotheses for all the failures—"not theories, just hypotheses." Why recruit? Maybe the guests are roostmates, who scout together. Maybe the birds fear ground predators and need numbers. Maybe they help each other tear at the meat or uncover it from the snow. Maybe they are kin anyhow, although ravens disperse hundreds of miles from their breeding nests to breed as solitary pairs defending their own territorial woods. Maybe... maybe.

Contemplation is not enough. It took a few tame ravens, if for nothing else than to check how to band them. It took building a forest cage big enough to trap a feeding crowd and band all of them. (The big birds were calm and dignified in the trap even though the graduate-student crowd recruited for the hard work of banding the fierce-beaked birds shouted from their shack with pleasure.) It took climbing 100-foot trees, to descend in utter exhaustion; it took freezing watches in blizzard weather; it took four years of hard work.

After four winters it was clear that ravens come quickly to recorded yells; that 90 percent of the eight tons of meat he manhandled to exposure sites was eaten by yelling crowds of ravens; that 90 percent of the crowd-feeders were juveniles; that those birds are highly vagrant, whereas paired adults are resident the year round; that pairs never yelled at the sight of meat; and that the lone juvenile discoverer does not yell until it has brought a few of its fellows to the scene quietly. "Conclusion: The recruitment is by vagrant juveniles" and mostly of juveniles. They thus gain access to food otherwise kept secret and defended by resident adult pairs. The crowd size does not depend on the size of the bait over a wide range, from 20 to 1,000 pounds. It is initial access that drives these acts more than final sharing. Probably the gregarious juveniles are led to recruit as a means of gaining status, demonstrating their fitness to mate. "It is an elegant, simple, and beautiful system."

The field notes here are uniformly evocative of life in the woods; perhaps they are too artless in their daily commonplaces to rank with the most readable natural history. But the explosive unfolding that marks the life of an investigation, with the explicit growth of evidence under sharp questioning, is brilliantly shared. His ravens advised Wotan; these ravens and their interpreter will inform any reader and no few philosophers of science. If only the ravens, too, could report their side of this long, unfinished encounter...

NATURAL GAS: BASIC SCIENCE AND TECHNOLOGY, by Alec Melvin. Adam Hilger, 1988. Distributed by Taylor & Francis, 242 Cherry Street, Philadelphia, PA 19106 ($75).

The thin fuelstuff hisses out of the wells by the gigaton each year, to make its intricate way along the finest tendrils of the pipelines, budding out in hundreds of millions of flames. "The free-burning flame... is perhaps too much taken for granted." The surprising fact is that we can easily stabilize a standing combustion wave only a fraction of a millimeter thick, a flame able to convert cool fuel and air to high-temperature products in a matter of milliseconds.

In this up-to-date review of the physics-based aspects of an industry that is somewhat inconspicuous for all its size and high technology, the longest chapter treats the working end of the gas stream: combustion and its products. The author, a physicist with wide responsibilities at British Gas, is sharply critical of current

progress in combustion science, seeing sterility and engineering irrelevance in the midst of a tide of "more papers than ever." No one said it would be easy. The simplest of flames that can be realized burns as a thin, flat disk across the smooth laminar flow of a carefully premixed stream of gas and air. Small, pointed flames of "neat fuel" must burn instead by the diffusion of outside air into their surface layers. Domestic flames are a hybrid of those two limits, whereas industrial furnaces harbor bigger, fiercer and fully turbulent flames.

Chemists see flame as something like a "high-temperature heat bath"; they use parameters derived from ideal laboratory flames to predict temperature profiles and stream composition. A typical numerical model of a premixed methane-oxygen flame includes 13 chemical species and 29 reactions! The fluid dynamicists are caught instead by the delicacy of stability conditions, which they study with simplified reaction models. This chapter pays most attention to particular models of flames and to experimental comparisons; its mathematical level goes beyond utility for a general reader even though most of this lucid and expert book stays within reach.

Flame behavior is qualitatively familiar enough. Stable flames burn a little way above the burner surface. Stream too slow, the flame will "light back" to ignite the incoming unburnt mixture ahead of the burner; stream too fast, the flame blows off and dies. Simplified models fit the measurements of spacing, temperature and composition pretty well, although not without some empirical and conceptual fudging. For the design of turbulent flames Melvin has hope that a more usable theory of turbulence may arise from the studies of chaos. The large computer programs widely used for the turbulent case are "literally uncriticizable by users and... require substantial faith."

Stable total combustion is the usual aim: energy efficiency. Today's central-heating boilers lose little heat up the flue, maybe 20 or 25 percent. A fan can reduce that loss to 18 percent, and with expensive heat exchangers only 5 percent need be wasted. Preheating improves heat yield, but there is a price to pay: the hotter the flame, the more nitrogen oxides it emits.

Big gas pipelines now lace Europe from Aberdeen and Barcelona far eastward past Sverdlovsk, and another web is spun over most of inhabited North America. The natural gases they transport at high pressure are typical-

1991 GERARD PIEL AWARD

FOR SERVICE TO SCIENCE IN THE CAUSE OF MAN

Nominations are requested for the fourth Gerard Piel Award for service to Science in the Cause of Man, to be presented by the International Council of Scientific Unions (ICSU) at its 23rd General Assembly in Sofia, Bulgaria, in October of 1990. The Award, established by the Board of Directors of Scientific American, Inc., was first bestowed on Gerard Piel, creator of the magazine *Scientific American*, upon his retirement as Chairman. The Award recognizes contributions to the wise use of science for the benefit of human welfare and fulfillment. It may recognize a lifelong or an episodic contribution to this cause. The prize will consist of a sum of $10,000 and a medal. Individuals and organizations are eligible. The Award is administered by a different scientific organization each year.

All nominations should include the following information, submitted on a typed letter: nominee's name, address, institutional affiliation and title; a brief biographical résumé, and a statement of justification for the nomination. Nominations of organizations should include information about the nature, form and work of the organization. All nominations must include the name, address, telephone number and signature of the person making the nomination.

Nominations, as well as questions about the Award, should be addressed to:

Executive Secretary
International Council of Scientific Unions
51 Blvd. de Montmorency
Paris 75016, France
Telephone: (33-1) 4525-0329
Telex: ICSU 630553 F
Telefax: (33-1) 4288-9431

Deadline for receipt of nominations is March 15, 1990.

ly 98 percent hydrocarbon fuels, although the mixture varies. Lean gases from Siberia may run only a couple of percent of the hydrocarbons heavier than methane, whereas rich ones, such as a North Sea gas, may contain 10 percent of ethane and propane.

The heat values of lean and rich gases of course differ. Less expected is the curious two-phase nature of the richer outflow. Over a certain range of pressures and temperatures, gaseous methane dissolves in the heavier fluid hydrocarbons. As the gas comes newly to the surface and loses its underground pressure, the flowing mix often separates into two phases. Such material flows through pipes in a variety of modes: as the pressure head changes along a vertical pipe, the stuff may move by as a bubbly liquid or, at the other end of the scale, as a spray of liquid droplets caught in a stream of gas. But it may take any of five other forms of two-phase flow, including a stratified gas-and-liquid mix, slugs of liquid within a flow of gas or occasional plugs of gas in a mostly liquid current. Compressor troubles, staccato vibrations, pipe erosion and general disorder may follow. The fluid dynamics is mostly empirical. Sometimes transport can be arranged at pressures high enough to hold the fuel always above the thermodynamic critical point of the two-phase mixture; that dense phase remains uniform.

Big gas volumes move under 75 atmospheres of pressure through steel pipes a yard wide at speeds of up to 40 or 50 miles per hour; big money flows with them. Where ownership changes along the way, people are eager to know the flow with precision and without delay. Errors of one part in 1,000 may imply millions of dollars of uncertainty in payment. The present-day scheme for such on-line flow measurements is almost clinical: ultrasound pulses travel back and forth between transducers within the pipe, and the sound transit times in the moving gas are closely monitored both with and against the stream. The mean flow speed is thus digitally sampled along a number of chords within the pipe; 200 pulses may be collected along each path every 10 seconds or so. This microprocessor-based scheme is so direct that "calibration" by conventional precision flowmeters checks not the new instrument but the standard one.

Other chapters here treat seismic-reflection prospecting, survey the entire operation (down to the Victorian piping sealed with lead and jute that still serves—leakily—many houses in London and Rome) and discuss in detail the thermodynamics of indirect measures of gas density.

An opening review looks at resources. By now the world has been reasonably well mapped; known conventional reserves of gas are widespread, the lion's share being Soviet, followed by Iran's. These reserves amount to under 100 gigatons, surely a lower limit. Gas remains more or less a local resource—at most a continental one; it does not travel well over the oceans.

Melvin's survey reminds us vividly of the possibilities—and the uncertainty—of a number of undeveloped gas resources, most of them larger by a good deal than the entire conventional reserve. Best known is the large amount of gas held so tightly in shale and sandstone that it invites artificial fracturing techniques; nuclear explosions have been tried by both superpowers to augment the flow.

Least certain, there may be outsize quantities of gas that never came from the decomposition of ancient life, as currently most gas and coal are held to have arisen. Most methane derives, a widely held Soviet theory argues, from reactions between carbon dioxide and water in the heat 100 miles down. A distinct case has been made (a book on the topic by Thomas Gold of Cambridge was reviewed in these pages in 1987) for primordial methane as old as the earth itself, still present in unprecedented quantity at much greater depths and steadily leaking upward. A single deep boring to meet it on the way has so far been only suggestive. Yet other types of gas deposit are known the world around, some too dilute, some too deep, some simply hard to extract. But natural gas is spillproof, and it is the least carbonaceous of fossil fuels. It is just what this polluted world longs to burn, if only we can get enough of it.

ALBUM OF SCIENCE: THE BIOLOGICAL SCIENCES IN THE TWENTIETH CENTURY, by Merriley Borell. **ALBUM OF SCIENCE: THE PHYSICAL SCIENCES IN THE TWENTIETH CENTURY,** by Owen Gingerich. Charles Scribner's Sons, 1989 ($75, each volume).

Five volumes of a treasury of scientific images are completed with this pair, a little ahead of the close of their epoch. I. Bernard Cohen, the scholar who conceived and as general editor guided the entire work, has with his companions reached his goal, a coherent display of the "pictorial record of the growth of the scientific enterprise" since antiquity. The frontispieces of these two volumes that share our century make eloquent markers. One is the first full model of the DNA double helix, viewed at the Cavendish Laboratory between two youthful investigators, Watson and Crick themselves; the other is an *Apollo* view of earthrise over the distant moon horizon, the lunar module just back in triumph from the Sea of Tranquillity and right in front of the lens.

It is photography that dominates the illustrations here, as it could not in any previous century. The witness of the artist, graphic representation by the scientists themselves, direct visual display of data and decisive settings in field, laboratory, museum, class and animal room are happily not forgotten. Both books, particularly that of the physical sciences, offer many images that must be called classic because they are so well known, a bittersweet experience for some readers but revelations for many.

Consider a random set of half a dozen classics chosen alternately from the two books: Freud's consultation room and its couch; Einstein, as young as Honest Jim, in his snappy checked suit of 1905; a sea-urchin egg with its bonded swarm of exigent sperm; the fogged plate from Henri Becquerel's drawer that first disclosed radioactive decay; the first electron micrograph of the feathery transcription of DNA; and the Great Red Spot (as well as Io and Miranda) from *Voyager 1.* (In these days of color plenty, our scholars appear in penury: not even one blue earth-marble here to certify present capabilities.)

Now for a few mint-fresh pictures delved from the "capricious archive of . . . history." The exponential growth of science and its literature characterizes the past few centuries; enter in vivid evidence the shot of an editor of *Physical Review* seated soberly between the piled issues of that journal for 1931, a set of papers hardly a foot high, and the same journal for 1985, bound into some 30 volumes, each as hefty as a metropolitan Yellow Pages. We see Linus Pauling in both books, once gravely next to his first of all models of a helical molecule crucial to life and again pacing cheerfully in shirtsleeves in front of the White House carrying his placard against atmospheric nuclear-weapons tests. The HeLa cultured cell line derives from tumor tissue removed from a feisty young black woman, seen here at a sunnier time. (The records do not preserve her name with surety; was it Helen Lane or Henrietta Lacks?)

The pioneer Naples Zoological Station, a white wedding cake outside and

plaster-white labs within, was the experimental "'battlefield where . . . zoological armies . . . fight . . . error and ignorance,'" its founder wrote. Two excavators stand well wrapped against the cold next to the ice-preserved foreleg and skull of the Berezovka River mammoth. The indigo research laboratory at BASF as our century opened is a columned Victorian hall, its chemists standing at their worktables in three-piece suits, some with hats on. The first detected action of a neutrino, at last something more than the passivity of unseen departure, is here in a complicated image. A suite of half a dozen frames from the history of germanium and silicon electronics carries the eye from the notebook entry of the discovery of the transistor effect past the point transistor, the first junction, the first integrated circuit and the first planar IC to VLSI and a glimpse of optical IC to come.

The biological volume is organized around issues, such as the shift away from the field toward an interventionist biology, human biology and medicine, our human place in nature . . . The physical volume is arranged more by subject: the atom, matter, the earth . . . The brief introductory texts that sum up each large division are often provocative (astrophysicist Gingerich has enlisted the help of specialist friends), although the main show is in the excellent and specific captions. (Picture sources are not given as fully as we might expect from historians; dates are usually missing.)

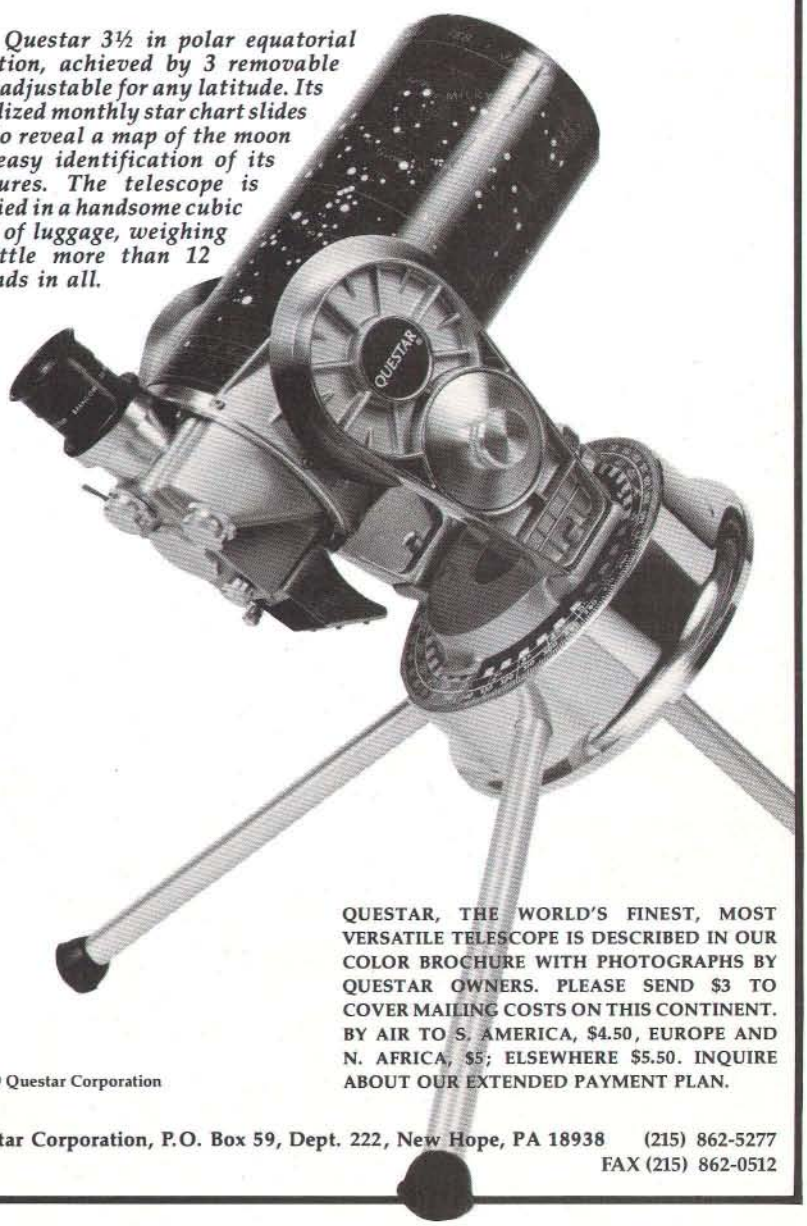
Both authors document the technological power that science has now conferred on our species and touch on the effect such power has had on our entire culture. They do not neglect to show us the dilemmas we face. It was cinematic imagination that built Fritz Lang's sleek, mutinous robot Maria; but it was real Agent Orange that ravaged the Vietnamese forest we see, and its impurities that may have sown widespread human birth defects. Turbulent mushroom clouds tower out of desert and lagoon. Every such event fills its frame quite well; that is the photographer's craft. But the camera stations were perforce at very different distances, a fact too often ignored, although explicit here. The thermonuclear specimen was a giant toxic mutant, even though it hardly looked different at the far-off lens.

These books, like their predecessors, offer both immediate and enduring value: for the concrete image is quickly grasped, and then an informed sense of wider pattern in the chronicle can slowly unfold.

THE USER-FRIENDLY
QUESTAR

In the past few years *user-friendly* has had quite a workout. In everything from computers to cameras we are being reassured: the gear that was hard to use is now easier; things are finally under our control. We don't feel we need to reassure you because the Questar telescope never was *user-unfriendly.* Instead we simply want to say that from the very start Questars have been designed for your pleasure, your comfort and your enjoyment. Lawrence Braymer's years of effort in perfecting this unique instrument were concerned equally with its quality and its ease of use. To bring optics and mechanics together in harmony was his first goal; to bring instrument and user into harmony was his second. As one example, the barrel revolves to put the eyepiece in the most comfortable position; another, the control box allows for instant changes of power. You can see that when we say user-friendly we really mean *user-delightful.* There's nothing like a Questar; why not let one please you?

The Questar 3½ in polar equatorial position, achieved by 3 removable legs adjustable for any latitude. Its anodized monthly star chart slides off to reveal a map of the moon for easy identification of its features. The telescope is carried in a handsome cubic foot of luggage, weighing a little more than 12 pounds in all.



QUESTAR, THE WORLD'S FINEST, MOST VERSATILE TELESCOPE IS DESCRIBED IN OUR COLOR BROCHURE WITH PHOTOGRAPHS BY QUESTAR OWNERS. PLEASE SEND $3 TO COVER MAILING COSTS ON THIS CONTINENT. BY AIR TO S. AMERICA, $4.50, EUROPE AND N. AFRICA, $5; ELSEWHERE $5.50. INQUIRE ABOUT OUR EXTENDED PAYMENT PLAN.

© 1989 Questar Corporation

Questar Corporation, P.O. Box 59, Dept. 222, New Hope, PA 18938 (215) 862-5277
FAX (215) 862-0512

ESSAY

Who will do science in the next century?

by Shirley M. Malcom

It has become almost a cliché to point out that U.S. science and technology are in trouble. The country is not training enough scientists and engineers, and that is largely because it is not convincing enough young people to opt for careers in science and engineering. Why? The finger is routinely pointed at science education in the elementary and secondary schools. The competing forces of cramming in (covering lots of facts), "dumbing down" (expecting little in the way of understanding) and weeding out (thinning the ranks) allow few students to emerge with their interest in science intact and with the kind of preparation that is essential for further education and careers in science or engineering.

For all of that, there exists a vast and strangely invisible talent pool that remains virtually untapped. Who are these people who would do science if they could? They are blacks and Hispanics and American Indians, girls and young women of all races, and disabled students of both sexes and all races. The great irony is that as bad as the educational system for science may be overall, its failure is most dismal precisely for members of these groups.

To maintain an internationally competitive science community and economy, the U.S. must meet the challenges of an array of converging demographic and educational trends.

The 18- to 24-year-olds who supply most of the new entrants into the work force and into colleges and universities are decreasing both in absolute number and as a proportion of the population.

Minority members make up a growing part of this shrinking cohort. By the year 2010 one in every three 18-year-olds will be black or Hispanic, in comparison with one in five in 1985.

Minority youngsters are or soon will be in the majority in the public schools of some states.

By the turn of the century minority members, women and immigrants will account for about 85 percent of net new members of the work force. Whereas only one generation ago women made up some 30 percent of the work force, by the year 2000 about half of the work force will be female.

If the U.S. economy keeps growing at its current rate, the need for scientists and engineers can be expected to increase or at least to remain the same. Even if it should stay the same, vacancies caused by attrition, by the peak of retirements expected in the late 1990's and by death will need to be filled.

The implication of these trends is clear. Unless we tap the students who are now in school and college—the very students who have not been going into science and engineering in large enough numbers—the U.S. will soon face a drastic shortage of young scientists. While the country tackles the long-term problem of restructuring elementary and secondary schools and their science and mathematics curricula, it therefore needs to face up to the short-term task: increasing the efficiency of universities in recruiting and graduating more Americans with degrees in science and engineering. In view of the changing demographic picture, "American" has simply got to mean minority and female as well as white and male.

That rather obvious lesson has not been learned. For example, counterproductive hiring and promotion practices by universities are slowing, and eventually may even reverse, the movement of women into graduate science education. Nearly 900 female citizens received doctorates in chemistry between 1978 and 1982. How is it that major research universities shunt so many of those women into peripheral positions as lecturers and instructors and admit so few of them to the tenure-track professorial ranks? As women students reject the notion that they may be good enough to train but not to hire, or to hire but not to promote, their dissatisfaction will likely feed back within the system, dampening supply.

Of the 341 citizens who received Ph.D. degrees in mathematics in 1988 (compared with 619 in 1978), only one was black, two were American Indians and three were Hispanic. Fewer than 80 black citizens have received doctorates in physics and astronomy in the past 11 years; fewer than 100 doctorates have been awarded to Hispanics in the same fields, and American Indians have averaged one a year. The trickle of minority scholars who leave the universities carrying a Ph.D. is so small that it will be difficult to diversify college faculties even as members of minorities come to make up more and more of the pool of college-age students.

Precisely because the numbers of minority science students are so small, small local efforts can change the numbers significantly. Walter E. Massey, chairman of the American Association for the Advancement of Science, has pointed out that if science departments of universities would commit to a "doubling plus one" of their current enrollment and Ph.D. output of such students, the results would make a difference immediately.

Yet the commitment to changing these statistics seems to be less fervent than the rhetoric. Science and engineering departments are still slow to offer graduate research assistantships to minority students. Such students may be accepted (particularly if they bring their own funding) but then may not be incorporated into the culture of the department. Without a mentor, feedback, a space to work, keys to the laboratory or access to colleagues, the student may be doomed to flounder. All too often, faculty members look at students who have sensory and mobility impairments and see the disability rather than the ability, in spite of such counterexamples as Stephen Hawking; too many male teachers look at female students and see helpers or potential conquests rather than talented, tenurable colleagues.

The opportunity to reverse these trends is at hand. Data suggest that between 1978 and 1988 interest in pursuing science and engineering majors grew among minority freshmen. In 1988, 11.5 percent of all black freshmen reported that they were interested in majoring in the physical sciences. We need to take them up on that readiness and to make sure that they overcome the 38 percent survival rate for all freshmen in those fields. Colleges tend to let the student kettle simmer and simply cream off the survivors; they need instead to develop and nurture talent whenever and wherever it can be identified. The next generation of science must necessarily draw on young people who are not generally seen (and indeed often do not see themselves) as being in the present mix.

Who will do science? That depends on who is included in the talent pool. The old rules do not work in the new reality. It's time for a different game plan that brings new players in off the bench.

SHIRLEY M. MALCOM is head of the Directorate for Education and Human Resources Programs of the American Association for the Advancement of Science.

Your business can reach the people who make the future happen in France.

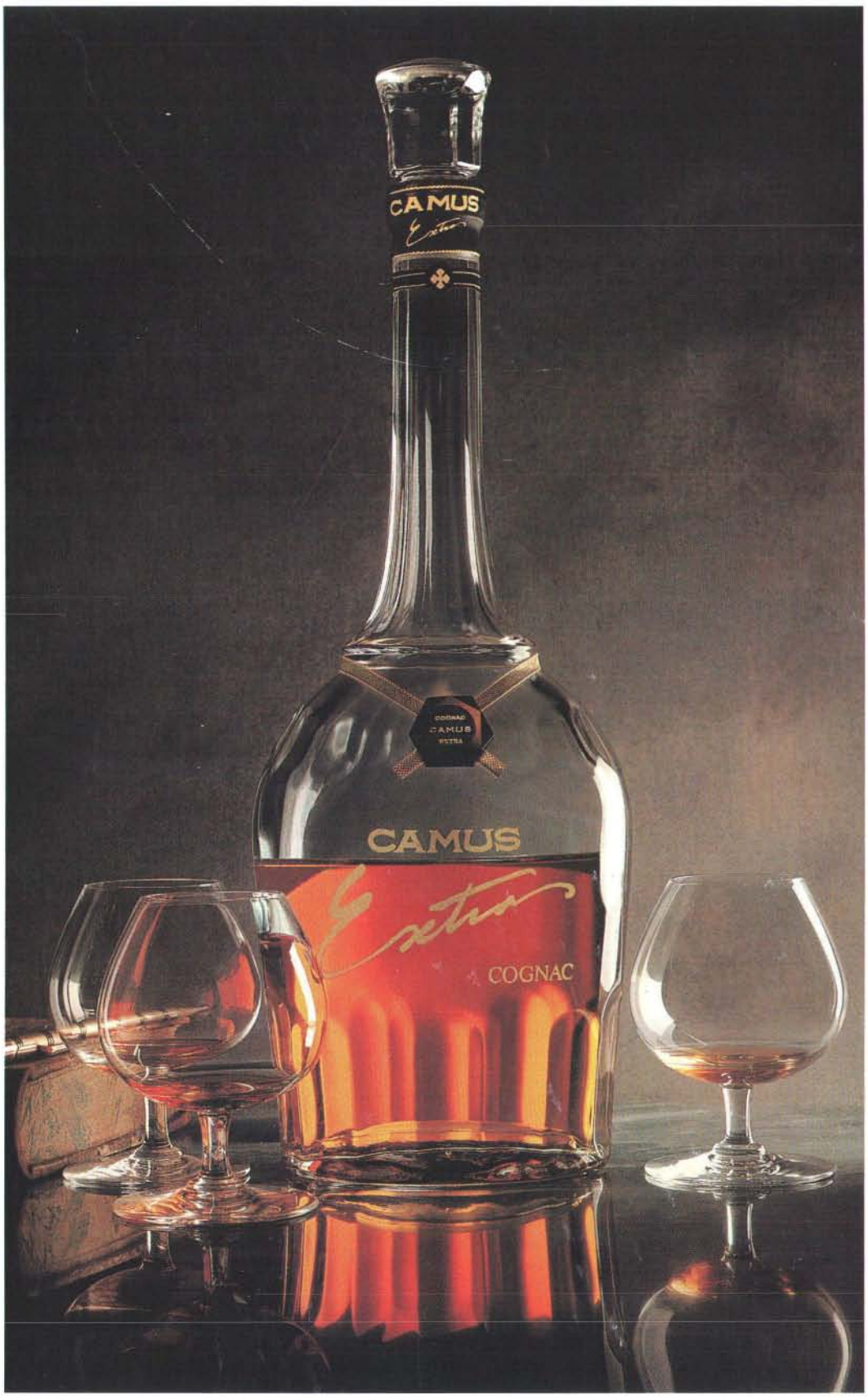
POUR LA SCIENCE is the French-language edition of SCIENTIFIC AMERICAN. More than 60,000 people buy this prestigious science and technology magazine each month, people who make decisions for France's industry, government and business sectors.

Contact
Susan Mackie at:

POUR LA SCIENCE
8, rue Férou
75006 Paris, France

Telephone
(33) (1) 46-34-21-42
Fax
(33) (1) 43-25-18-29
Telex
842202978

Clearly the judges had no difficulty in voting Camus the best cognac in the world.



In 1984, we at Camus decided for the first time to enter our XO Cognac in the International Wine and Spirits Competition.

Camus XO was deliberated upon by a collection of the most highly-qualified palates in the world, who duly pronounced the Camus XO a worthy winner of the gold medal.
In 1987, we entered again, this time with Camus Extra.

Not surprisingly it, too, won the gold medal, leaving Camus with the enviable record of two entries and two gold medals.
Incidentally, no gold award was given in 1988. Coincidentally, Camus did not enter that year.

CAMUS
COGNAC, FRANCE